← Speaker: Rodrigo Benenson

Weakly supervised semantic segmentation

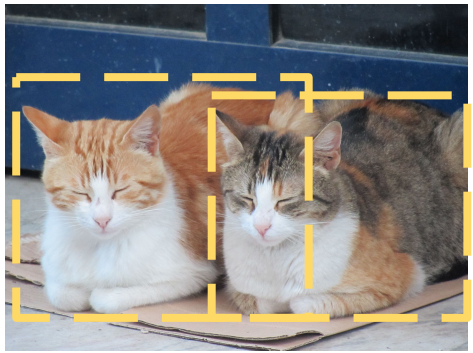
Classification

Is there a cat?

Classification

Is there a cat?

Detection

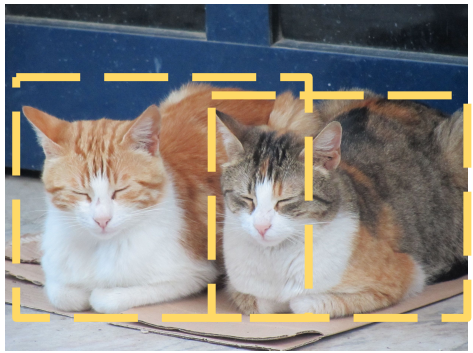


Where is the cat?

Classification

Is there a cat?

Detection



Where is the cat?

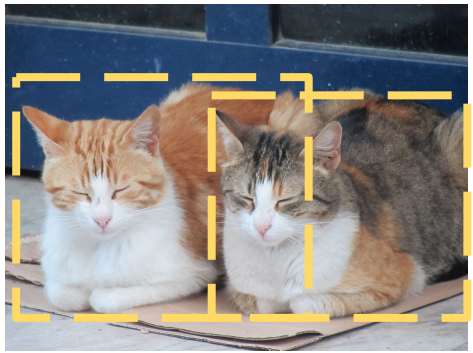
Semantic labeling

Which pixel is cat ?

Classification

Is there a cat?

Detection



Where is the cat?

Semantic labeling

Which pixel is cat ?

Instance segmentation

Which are the cat’s pixels?

Classification

Is there a cat?

Detection

Where is the cat?

Semantic labeling

Which pixel is cat ?

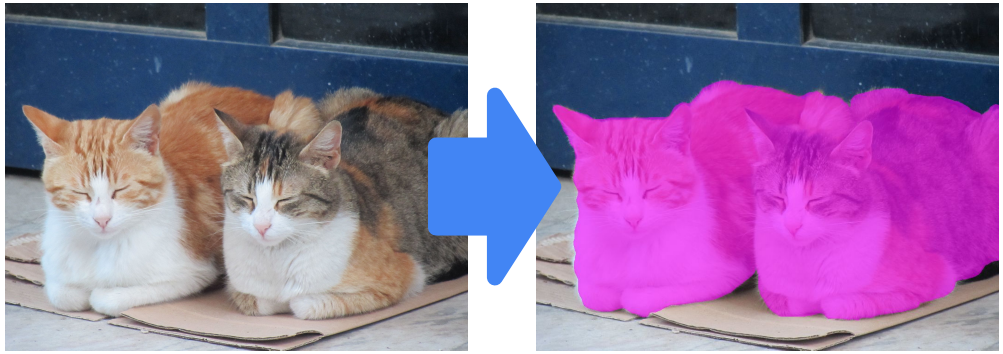
Instance segmentation

Which are the cat’s pixels?

+5 papers

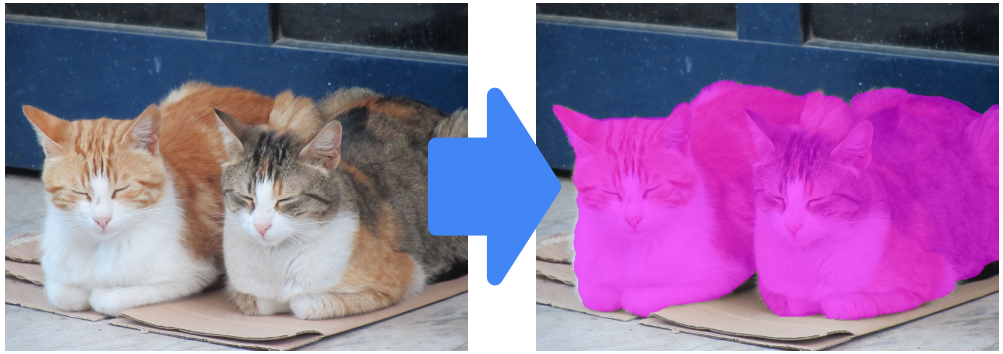
KEEP
CALM
AND
WRITE DOWN
QUESTIONS

Fully supervised → **Directly** supervised



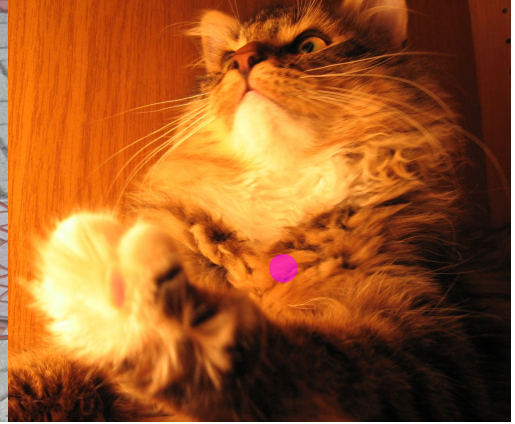
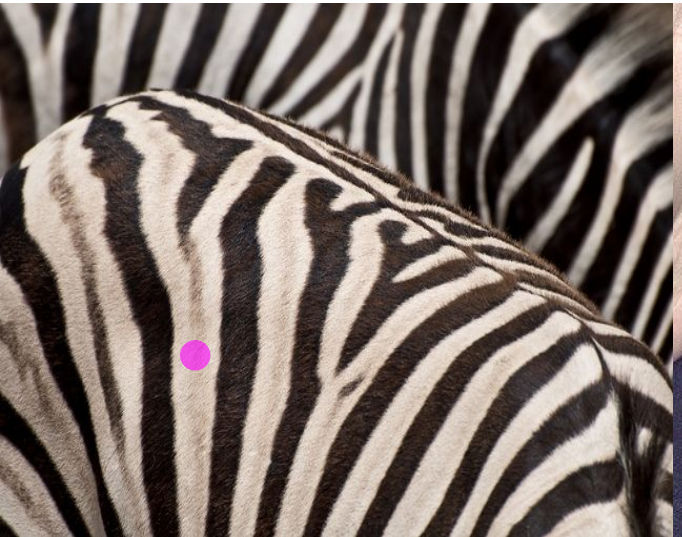
For each input, the desired output is provided

Fully supervised → **Directly** supervised



Weakly supervised → **Indirectly** supervised



Maker's
Mark
WHISKY


There are only two sources of information:

Priors & Hints

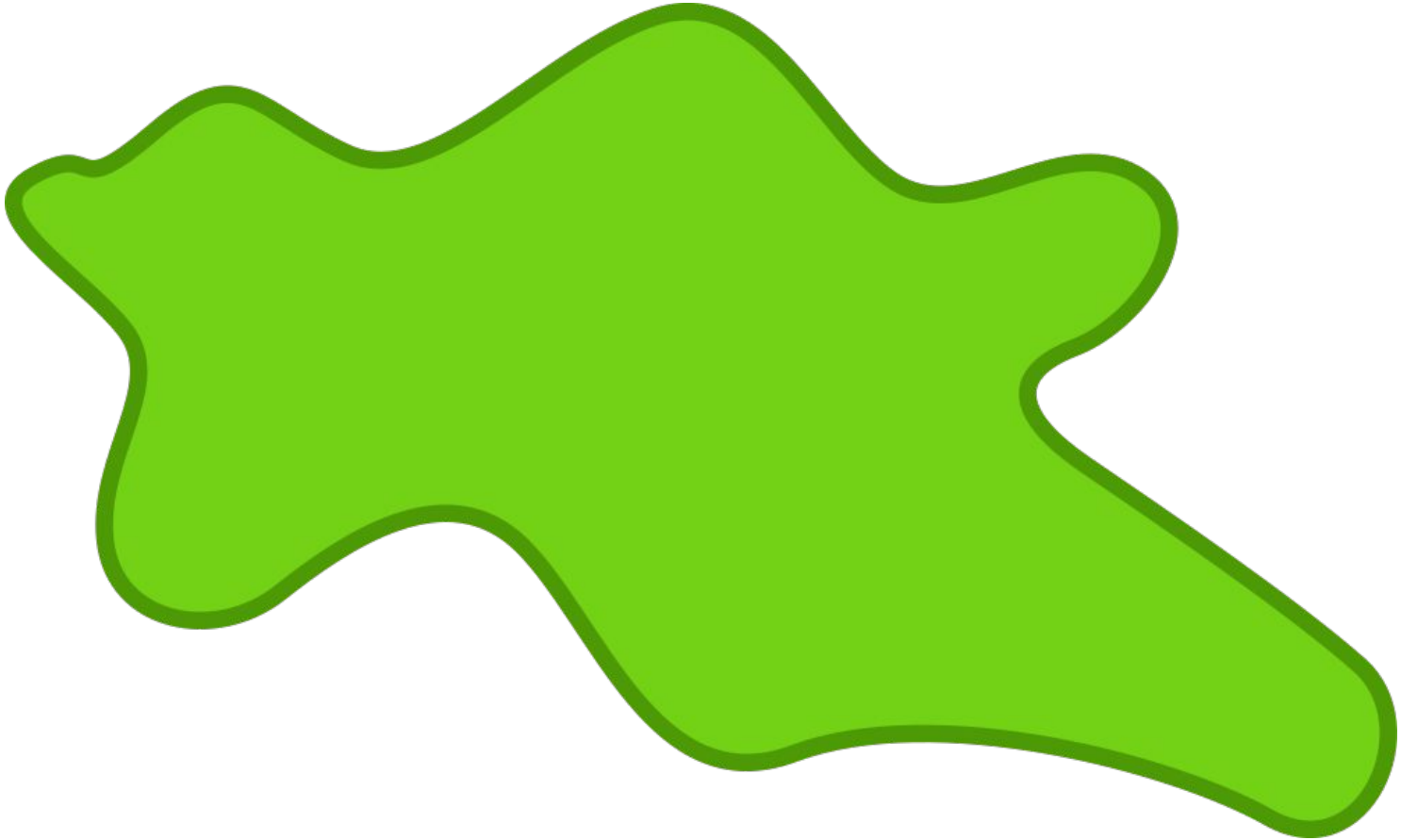
Priors & Hints

Priors: what do I believe to be true independent of any particular image sample

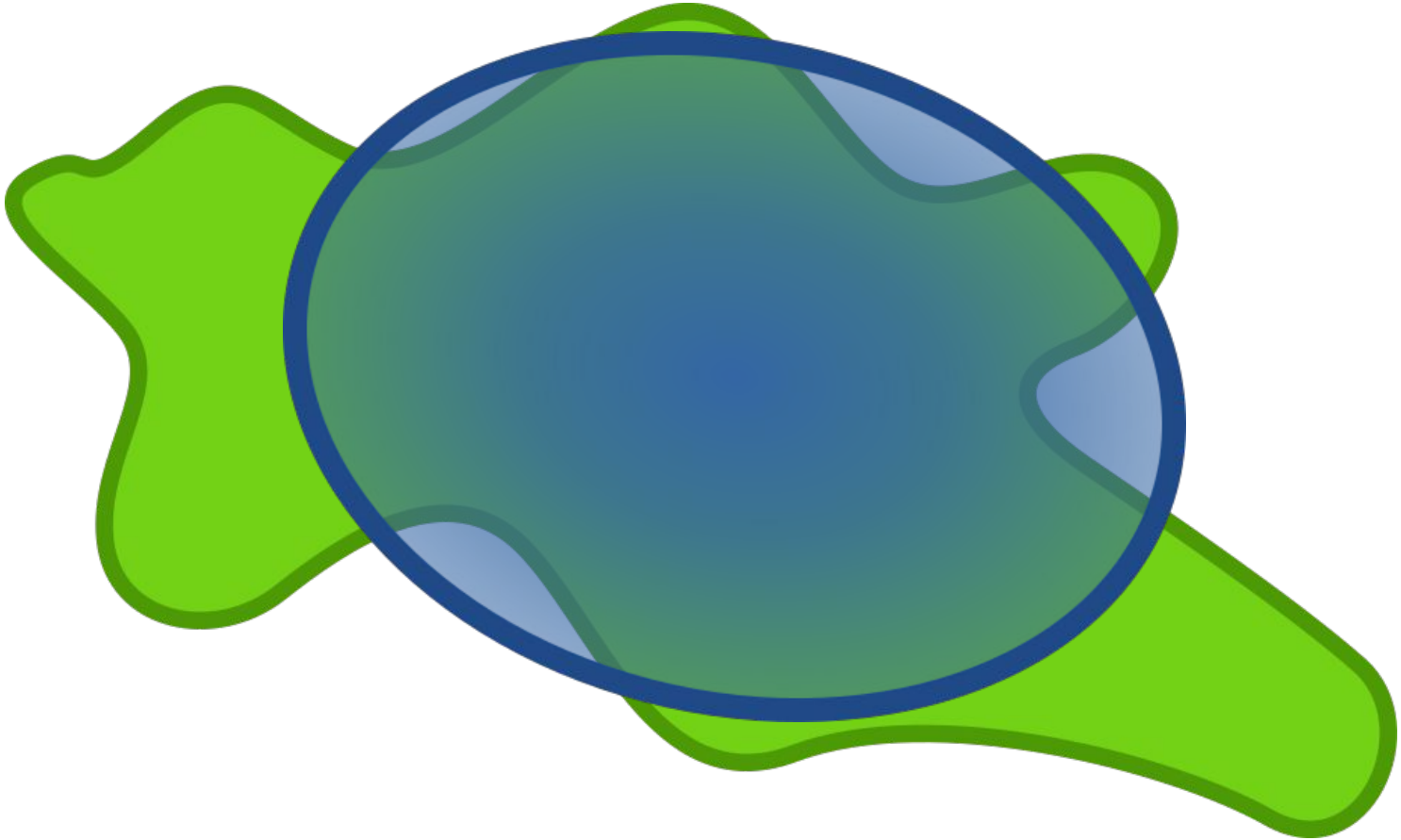
Priors & Hints

Hints: indirect supervision received for each image

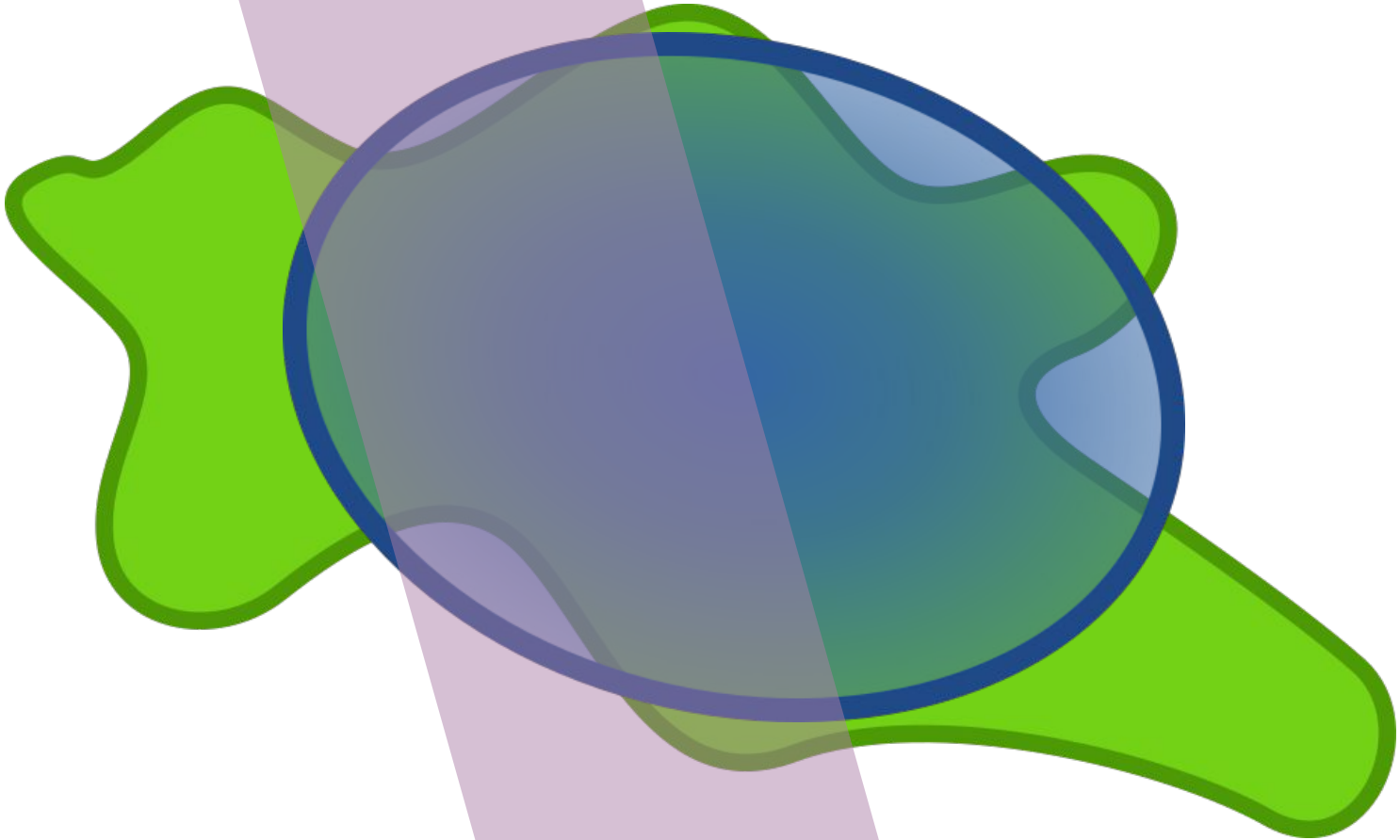
Priors & Hints provide constraints



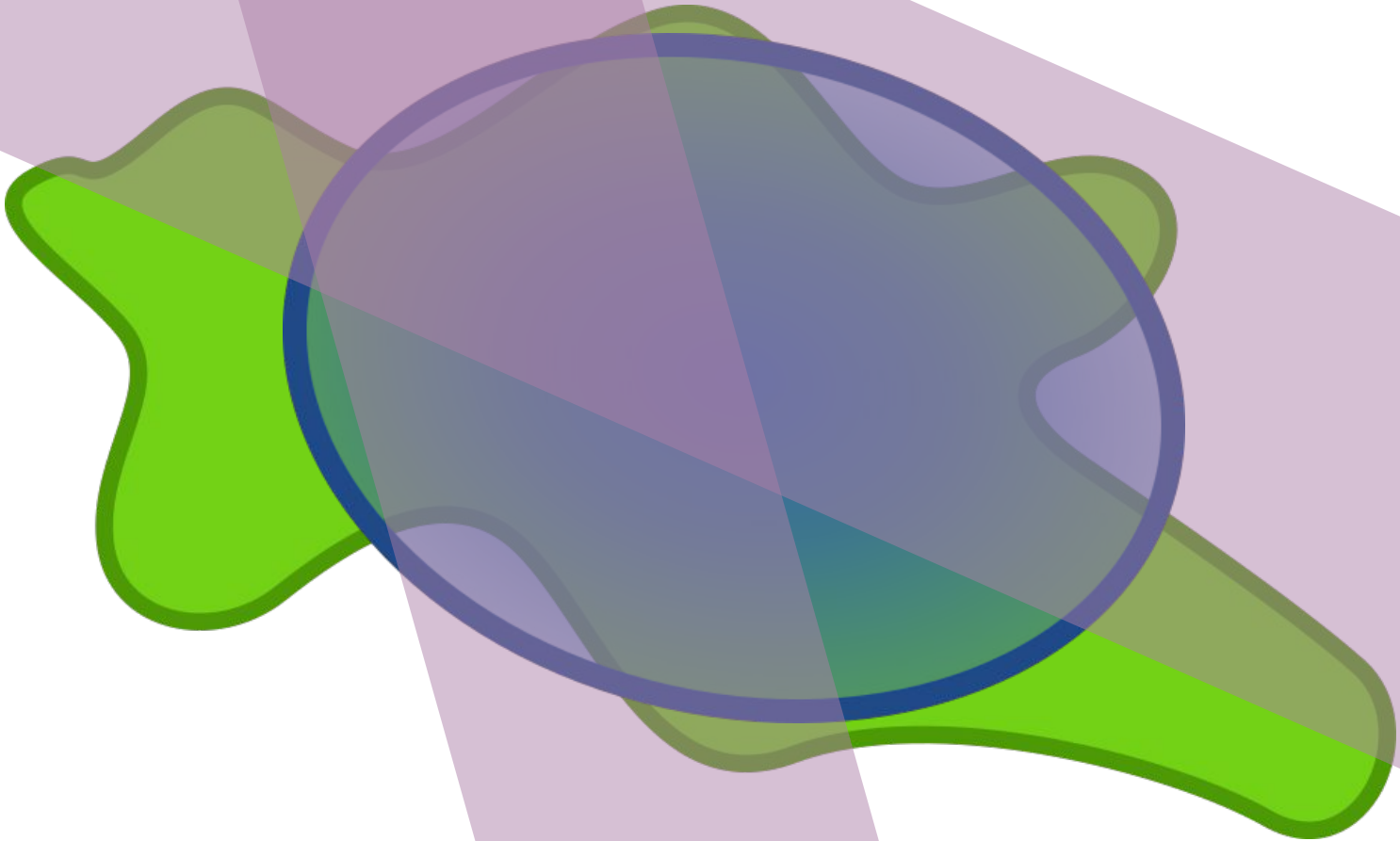
Priors & Hints provide constraints



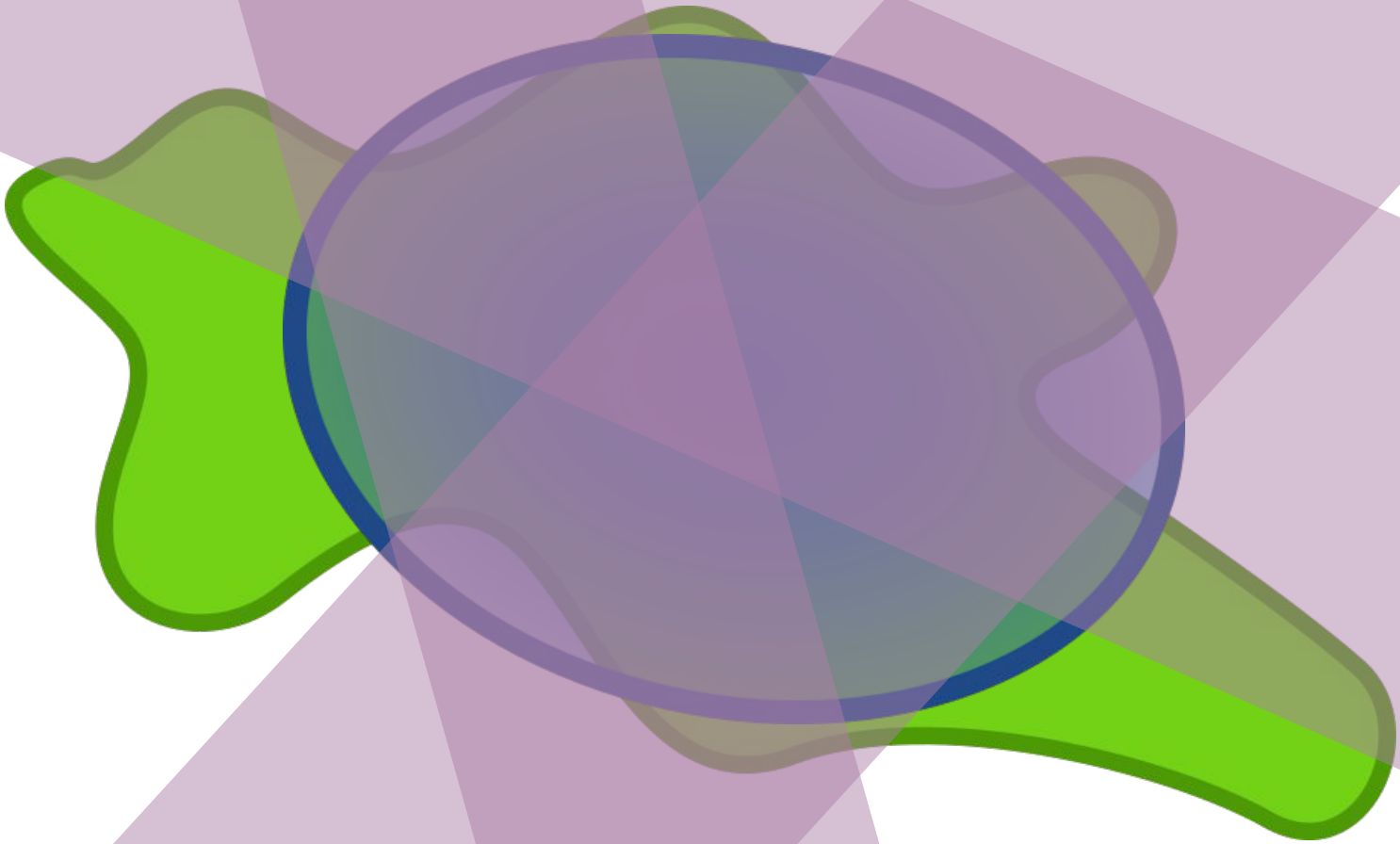
Priors & Hints provide constraints



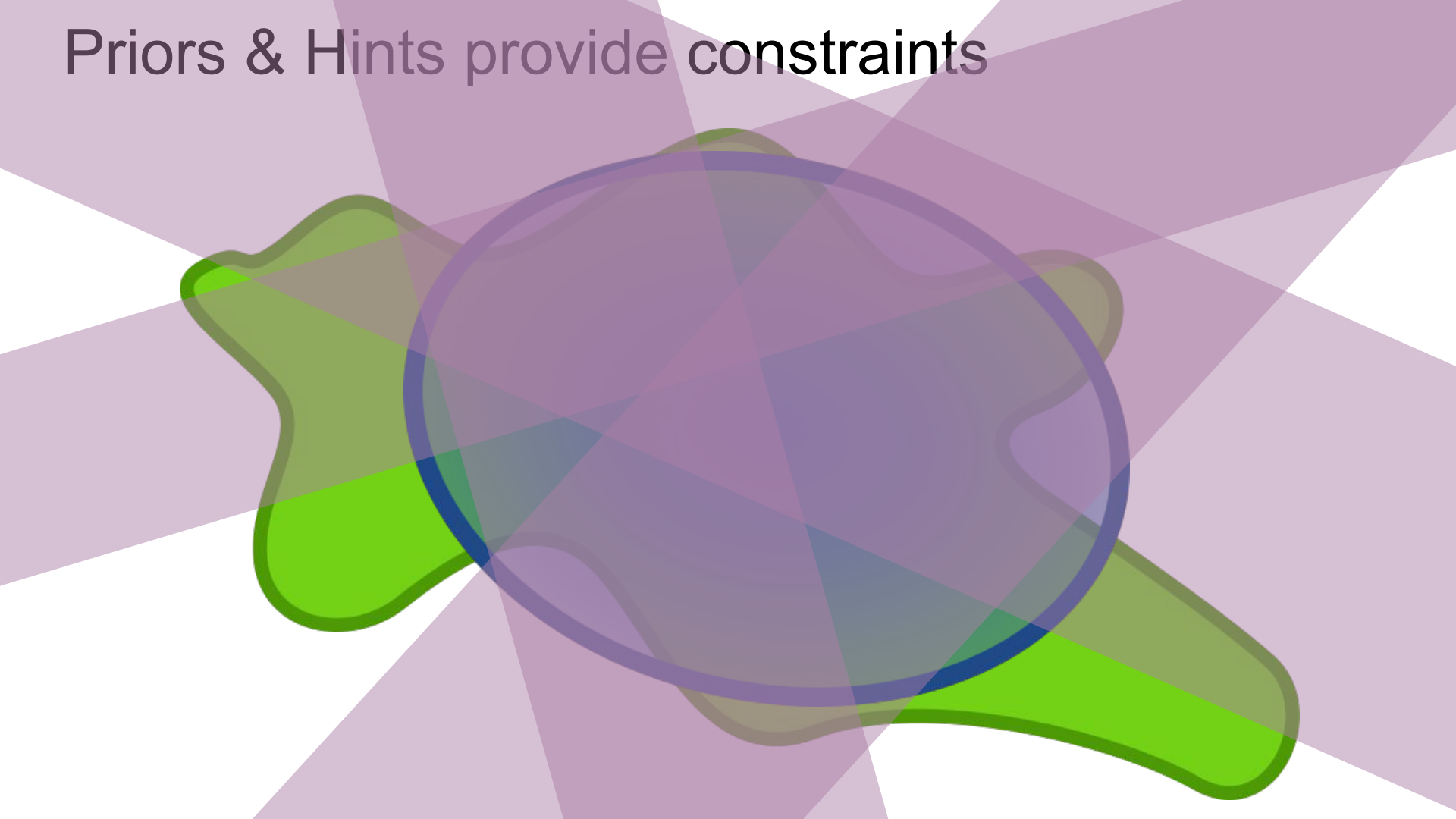
Priors & Hints provide constraints



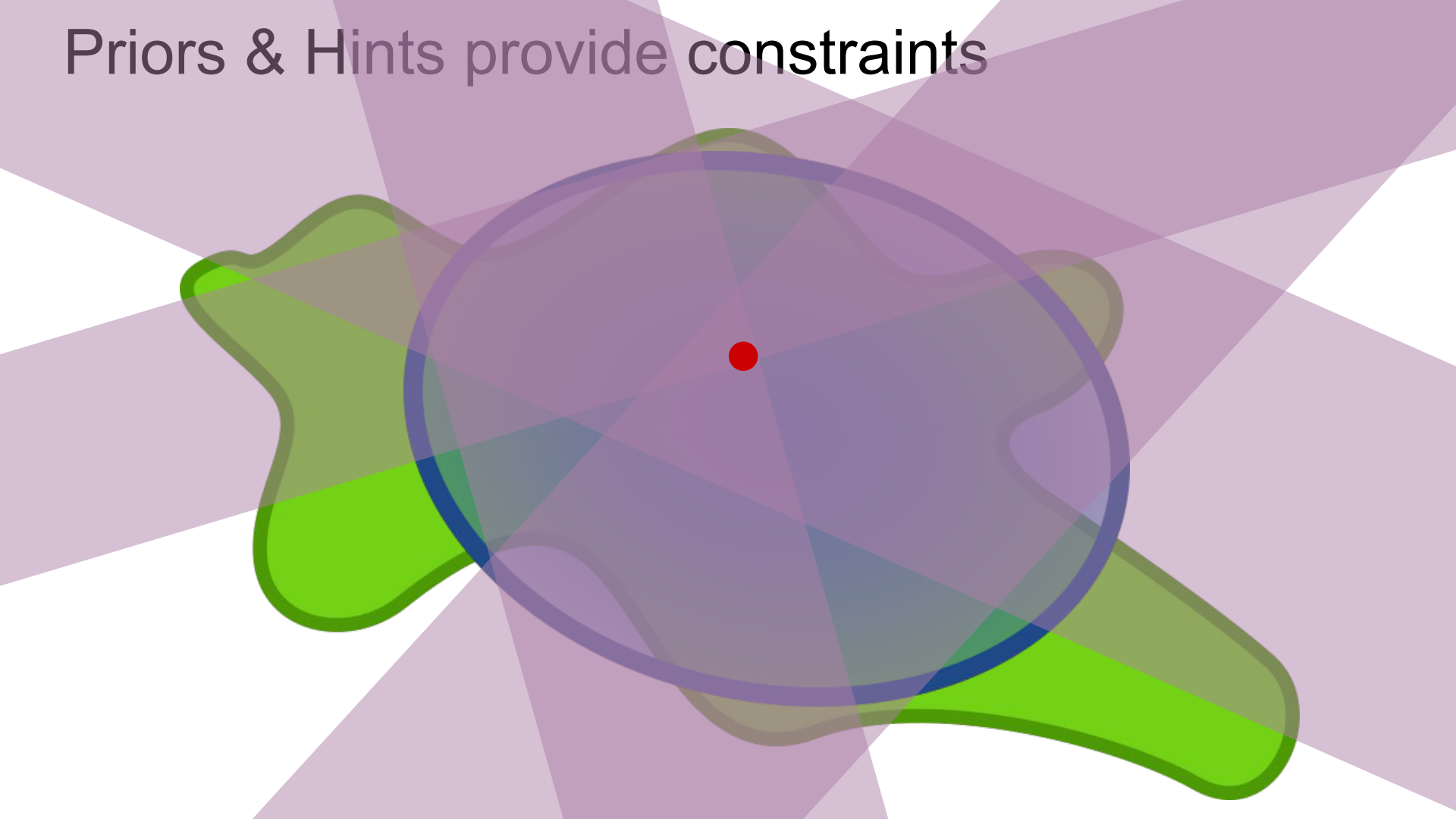
Priors & Hints provide constraints



Priors & Hints provide constraints

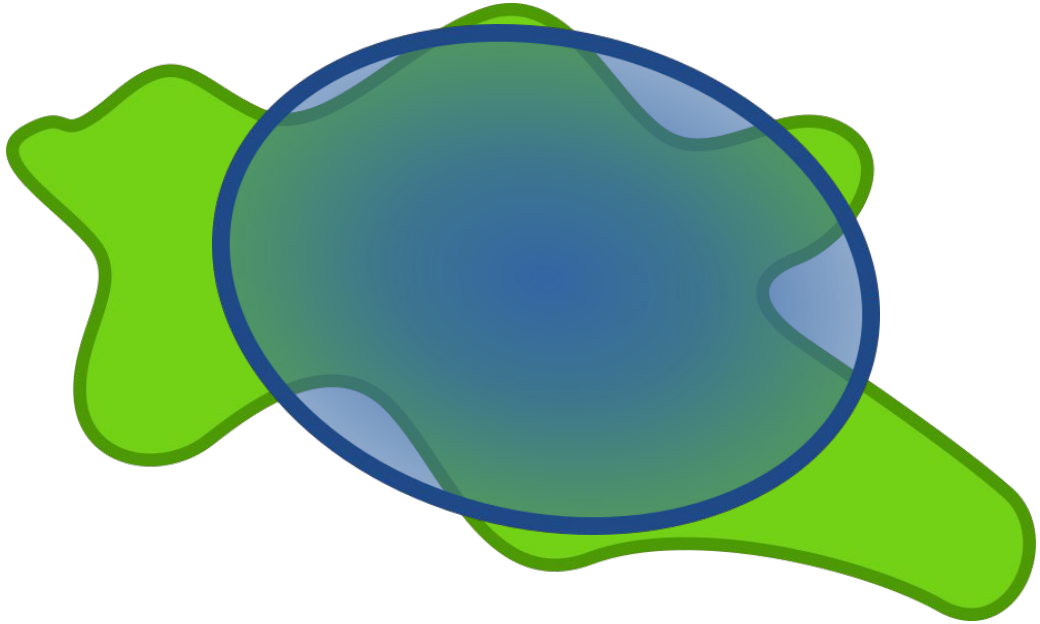


Priors & Hints provide constraints



Priors

- Size
- Shape
- Location
- Number of instances
- Contrast (boundaries, saliency)
- Class distribution
- Motion
- Similarity across images
- Similarity with external images



Priors

- **Size**
- Shape
- Location
- Number of instances
- Contrast (boundaries, saliency)
- Class distribution
- Motion
- Similarity across images
- Similarity with external images

Priors

- **Size**
- Shape
- Location
- Number of instances
- Contrast (boundaries, saliency)
- Class distribution
- Motion
- Similarity across images
- Similarity with external images



versus



[Pathak ICCV 15, Xu CVPR 15]

Priors

- Size
- **Shape**
- Location
- Number of instances
- Contrast (boundaries, saliency)
- Class distribution
- Motion
- Similarity across images
- Similarity with external images



versus



[Gulshan CVPR 10, Tang ICCV 13]

Priors

- Size
- Shape
- **Location**
- Number of instances
- Contrast (boundaries, saliency)
- Class distribution
- Motion
- Similarity across images
- Similarity with external images



versus



[most papers implicitly, e.g. Remez arxiv 18 explicitly]

Priors

- Size
- Shape
- **Location**
- Number of instances
- Contrast (boundaries, saliency)
- Class distribution
- Motion
- Similarity across images
- Similarity with external images



versus



[most papers implicitly, e.g. Remez ECCV 18 explicitly]

Priors

- Size
- Shape
- Location
- **Number of instances**
- Contrast (boundaries, saliency)
- Class distribution
- Motion
- Similarity across images
- Similarity with external images



versus



[most WSD papers implicitly, e.g. Deselaers ECCV 10]

Priors

- Size
- Shape
- Location
- Number of instances
- **Contrast** (boundaries, saliency)
- Class distribution
- Motion
- Similarity across images
- Similarity with external images

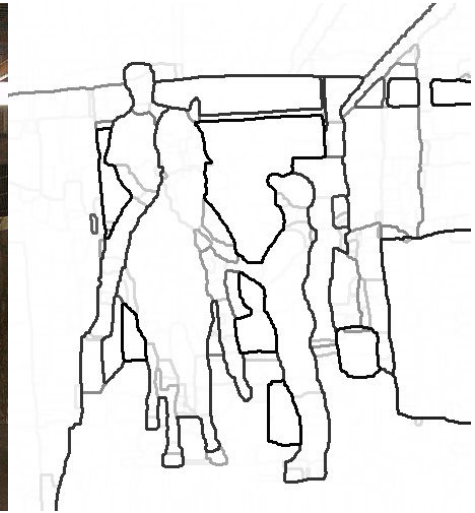


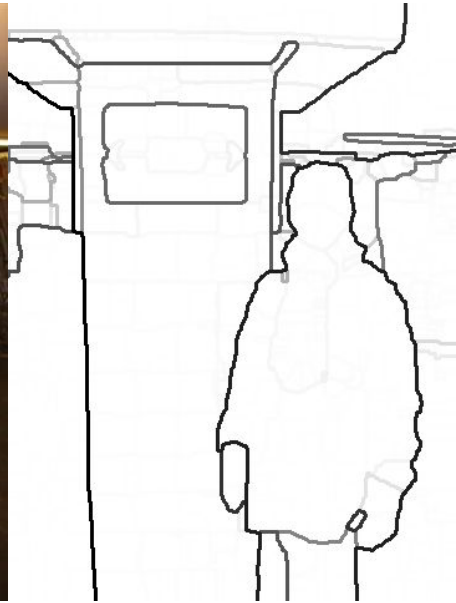
versus



[all WSS papers]

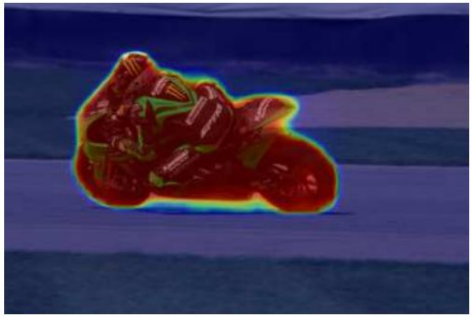
Contrast: **boundaries;**
Which pixels separate objects ?





[Maninis ECCV 16]

Contrast: **saliency**;
What is the subject of the photo ?



[Cheng PAMI 15, Oh CVPR 17]

Priors

- Size
- Shape
- Location
- Number of instances
- Contrast (boundaries, saliency)
- Class distribution
- **Motion**
- Similarity across images
- Similarity with external images

[Tokmakov ECCV 16, Huang ECCVw 17]

Priors

- Size
- Shape
- Location
- Number of instances
- Contrast (boundaries, saliency)
- **Class distribution**
- Motion
- Similarity across images
- Similarity with external images



⅔ Cats, ⅓ Cars

[Pathak ICCV 15, Xu CVPR 15]

Priors

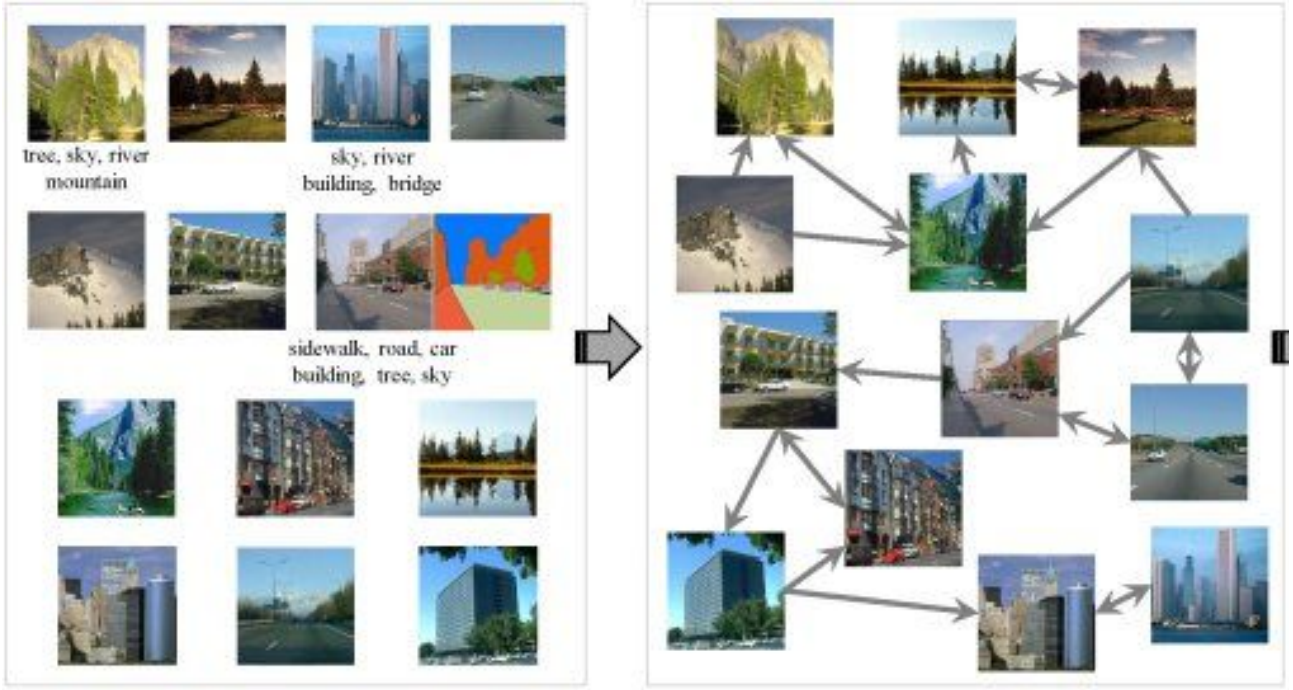
- Size
- Shape
- Location
- Number of instances
- Contrast (boundaries, saliency)
- Class distribution
- Motion
- **Similarity across images**
- Similarity with external images

versus

[all WSS papers, e.g. Sun et al. 2020]

Priors

- Size
- Shape
- Location
- Number of instances
- Contrast (boundaries, saliency)
- Class distribution
- Motion
- Similarity across images
- **Similarity with external images**



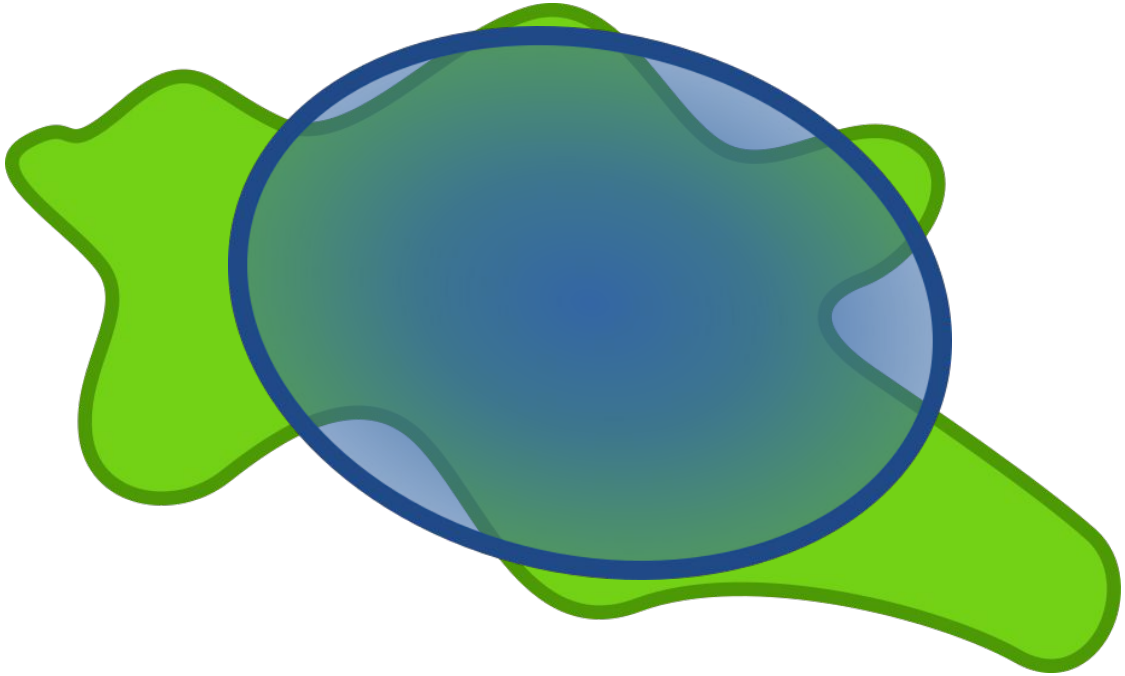
(a) Input database + sparse annotations

(b) Annotation propagation
(→ = dense pixel correspondence)

[Kuettel ECCV 12, Rubinstein ECCV 12]

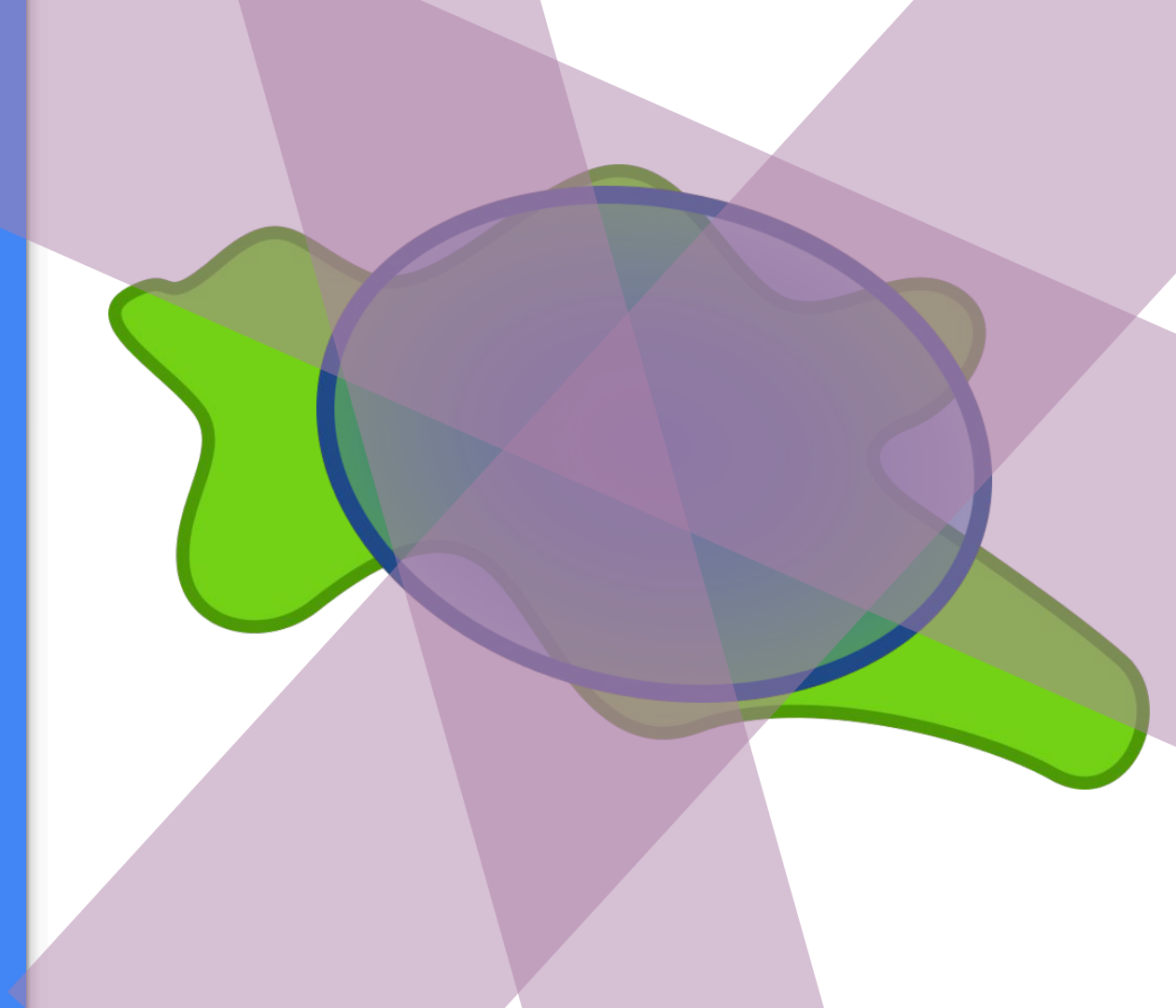
Priors

- Size
- Shape
- Location
- Number of instances
- Contrast (boundaries, saliency)
- Class distribution
- Motion
- Similarity across images
- Similarity with external images



Hints

- Image labels
- Image captions
- Video labels
- Transfer across images
- Transfer across classes
- Click inside object
- Size-from-center-click
- Object bounding boxes
- Objects extreme points
- Scribbles
- Eye gaze
- Localized narratives



Hints

- **Image labels**
- Image captions
- Video labels
- Transfer across images
- Transfer across classes
- Click inside object
- Size-from-center-click
- Object bounding boxes
- Objects extreme points
- Scribbles
- Eye gaze
- Localized narratives

Person, cat, door

[Verbeek CVPR 07, Vezhnevets ICCV 11, Xu CVPR 14, Pinheiro CVPR 15, Pathak ICLR 15, Papandreou ICCV 15, Kolesnikov ECCV 16, Wei CVPR 17, Singh ICCV 17]

Hints

- Image labels
- **Image captions**
- Video labels
- Transfer across images
- Transfer across classes
- Click inside object
- Size-from-center-click
- Object bounding boxes
- Objects extreme points
- Scribbles
- Eye gaze
- Localized narratives

Kattie holds the cat next to the door

[Rohrbach ECCV 16, Hu ECCV 16]

Hints

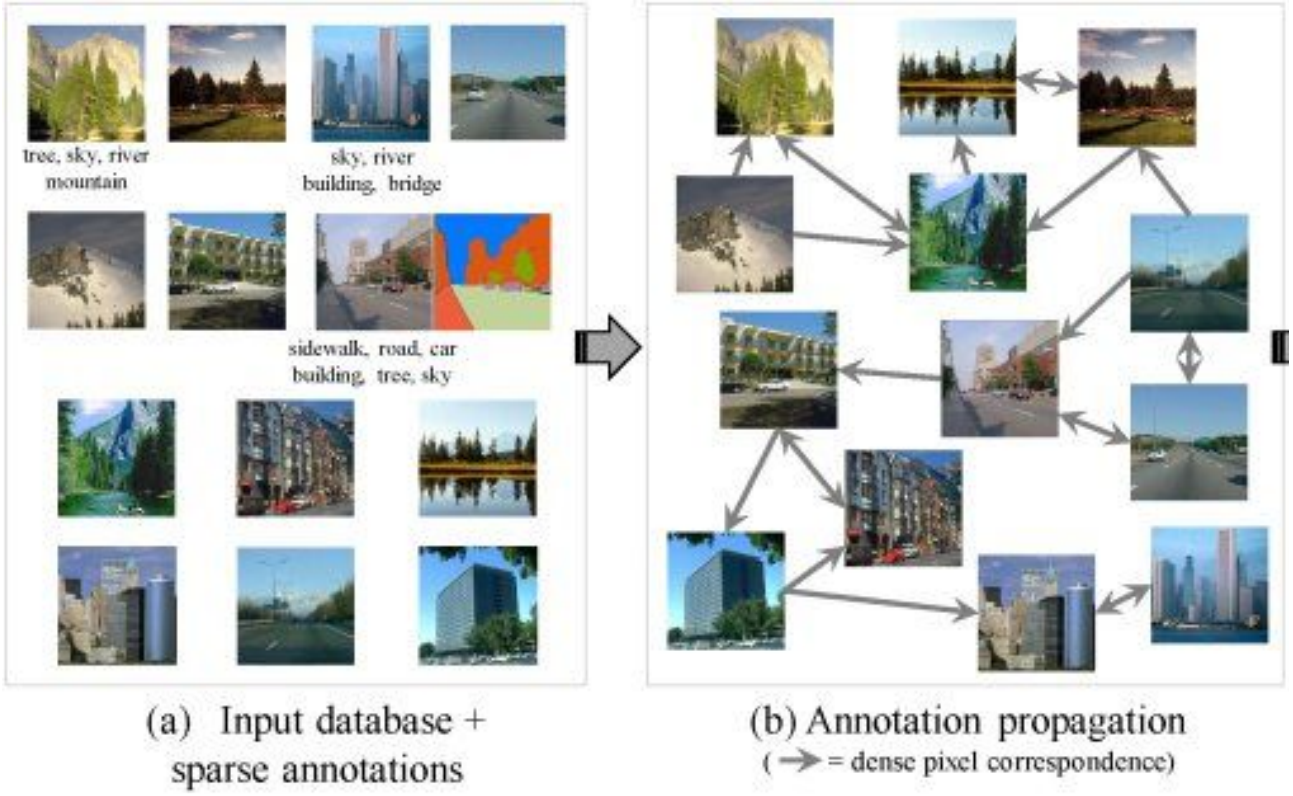
- Image labels
- Image captions
- **Video labels**
- Transfer across images
- Transfer across classes
- Click inside object
- Size-from-center-click
- Object bounding boxes
- Objects extreme points
- Scribbles
- Eye gaze
- Localized narratives

Cat, Curtain, Chair

[Tokmakov ECCV 16]

Hints

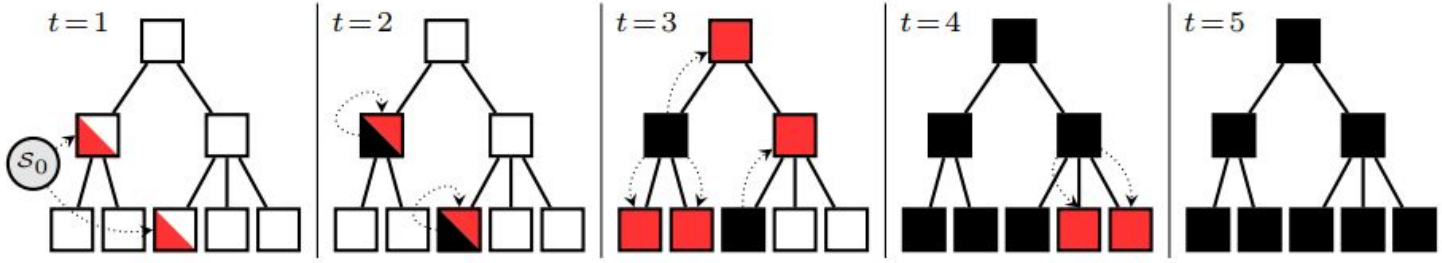
- Image labels
- Image captions
- Video labels
- **Transfer across images**
- Transfer across classes
- Click inside object
- Size-from-center-click
- Object bounding boxes
- Objects extreme points
- Scribbles
- Eye gaze
- Localized narratives

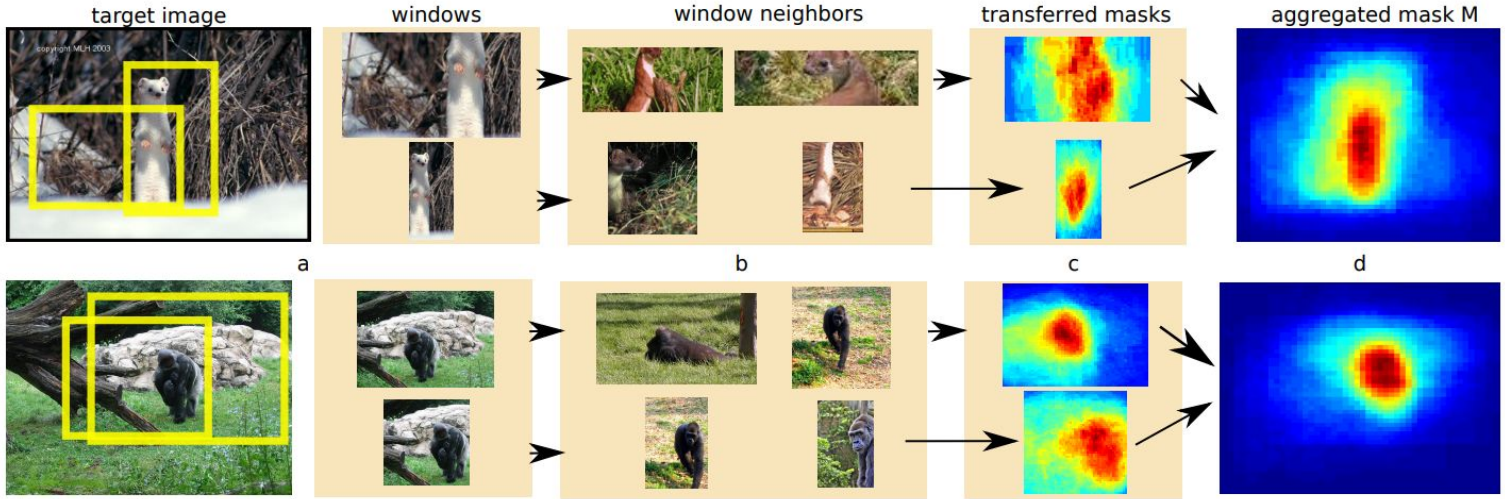


[Rubinstein ECCV 12]

Hints

- Image labels
- Image captions
- Video labels
- Transfer across images
- **Transfer across classes**
- Click inside object
- Size-from-center-click
- Object bounding boxes
- Objects extreme points
- Scribbles
- Eye gaze
- Localized narratives

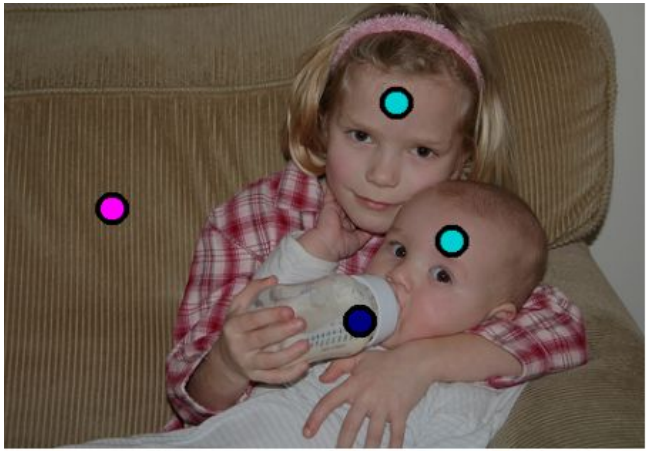




[Kuettel ECCV 12, Lim NIPS 11]

Hints

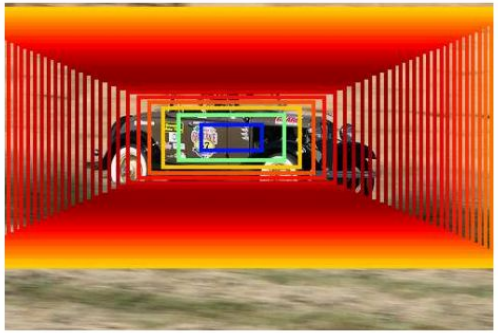
- Image labels
- Image captions
- Video labels
- Transfer across images
- **Click inside object**
- Size-from-center-click
- Object bounding boxes
- Objects extreme points
- Scribbles
- Eye gaze
- Localized narratives



[Wang CVIU 14, Bell CVPR 15, Bearman ECCV 16, Jain AAAI 16]

Hints

- Image labels
- Image captions
- Video labels
- Transfer across images
- Click inside object
- **Size-from-center-click**
- Object bounding boxes
- Objects extreme points
- Scribbles
- Eye gaze
- Localized narratives



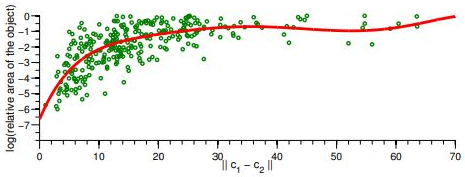
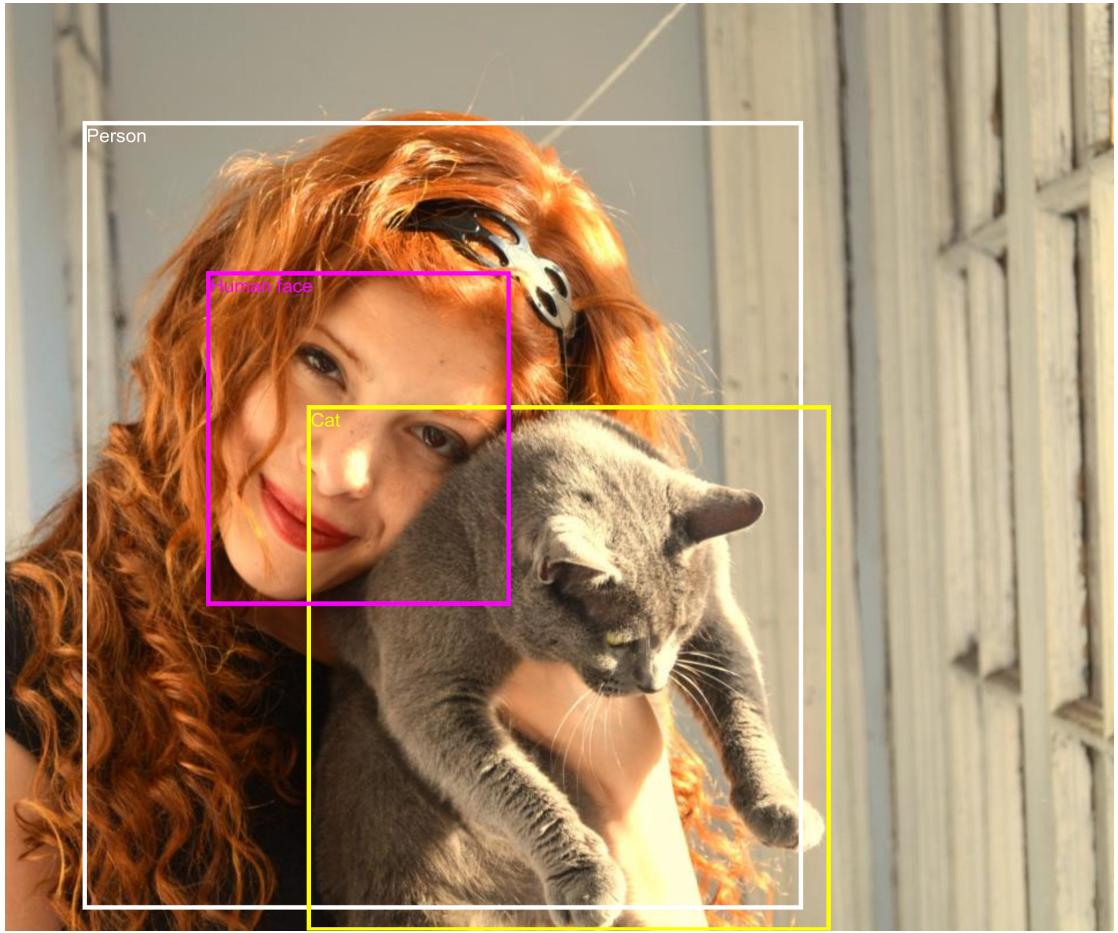


Figure 6. **Box area score** S_{ba}. All windows used here have fixed aspect ratio and are centered on the center of the object.

[Papadopoulos CVPR 17]

Hints

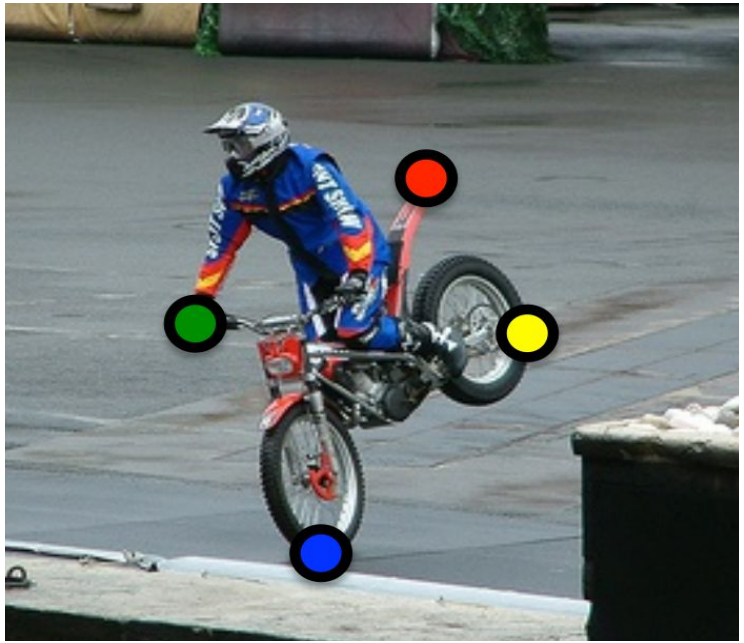
- Image labels
- Image captions
- Video labels
- Transfer across images
- Click inside object
- Size-from-center-click
- **Object bounding boxes**
- Objects extreme points
- Scribbles
- Eye gaze
- Localized narratives

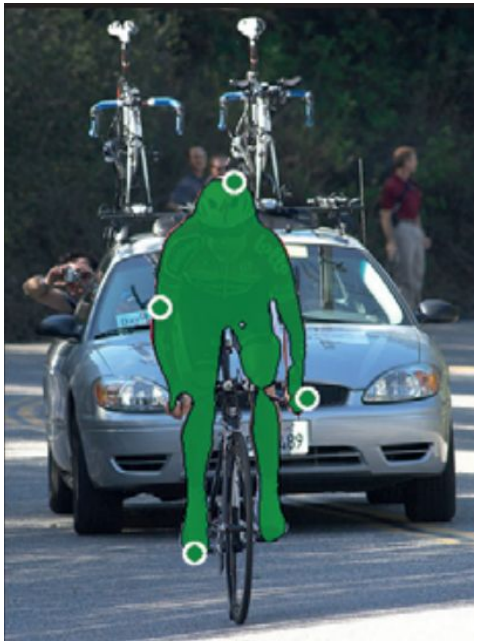


[Rother SIGGRAPH 04, Dai ICCV 15, Khoreva CVPR 17]

Hints

- Image labels
- Image captions
- Video labels
- Transfer across images
- Click inside object
- Size-from-center-click
- Object bounding boxes
- **Objects extreme points**
- Scribbles
- Eye gaze
- Localized narratives

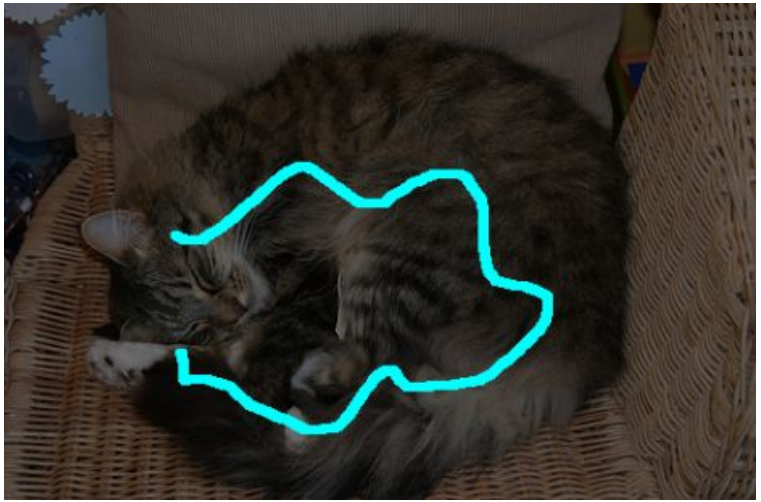




[Papadopoulos ICCV 17, Maninis CVPR 18]

Hints

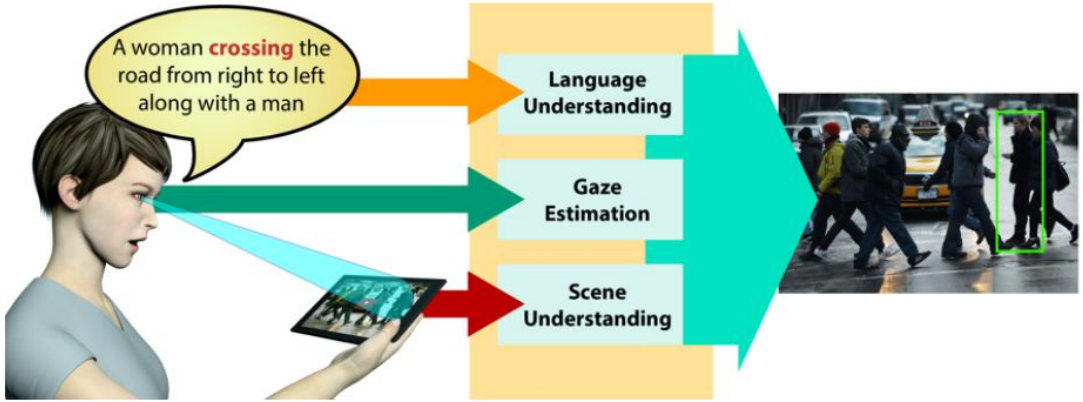
- Image labels
- Image captions
- Video labels
- Transfer across images
- Click inside object
- Size-from-center-click
- Object bounding boxes
- Objects extreme points
- **Scribbles**
- Eye gaze
- Localized narratives



[Xu CVPR 15, Lin CVPR 16, Bearman ECCV 16]

Hints

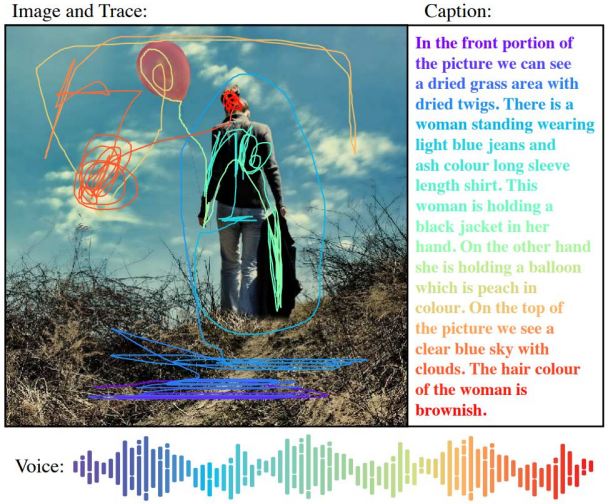
- Image labels
- Image captions
- Video labels
- Transfer across images
- Click inside object
- Size-from-center-click
- Object bounding boxes
- Objects extreme points
- Scribbles
- **Eye gaze**
- Localized narratives

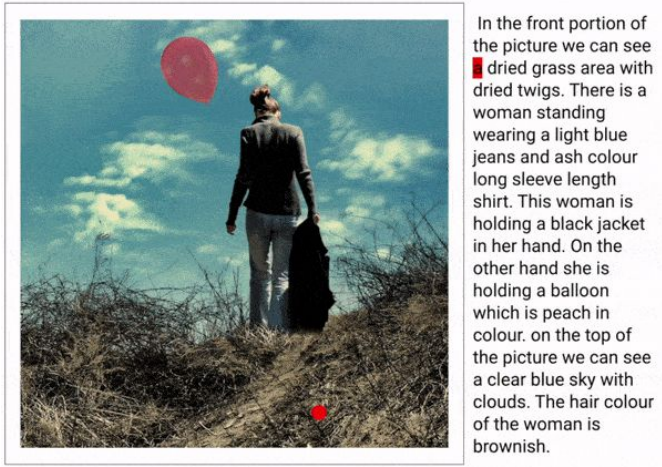


[Vasudevan CVPR 18]

Hints

- Image labels
- Image captions
- Video labels
- Transfer across images
- Click inside object
- Size-from-center-click
- Object bounding boxes
- Objects extreme points
- Scribbles
- Eye gaze
- **Localized narratives**



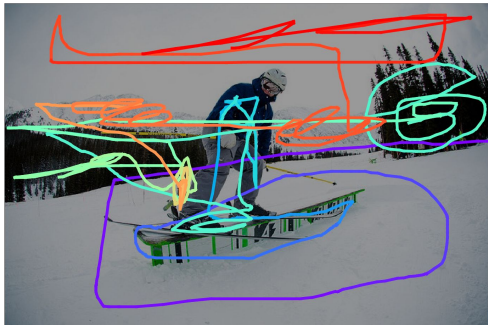


"A man skiing along a rail on the snowy hill."

Standard Caption

"In the foreground of the picture there is snow and a railing. On the railing there is a person in blue dress skiing. On the right there are trees. On the left there are trees. Behind the man there is another person in red dress walking. In the center there are mountains. On the top sky is cloudy."

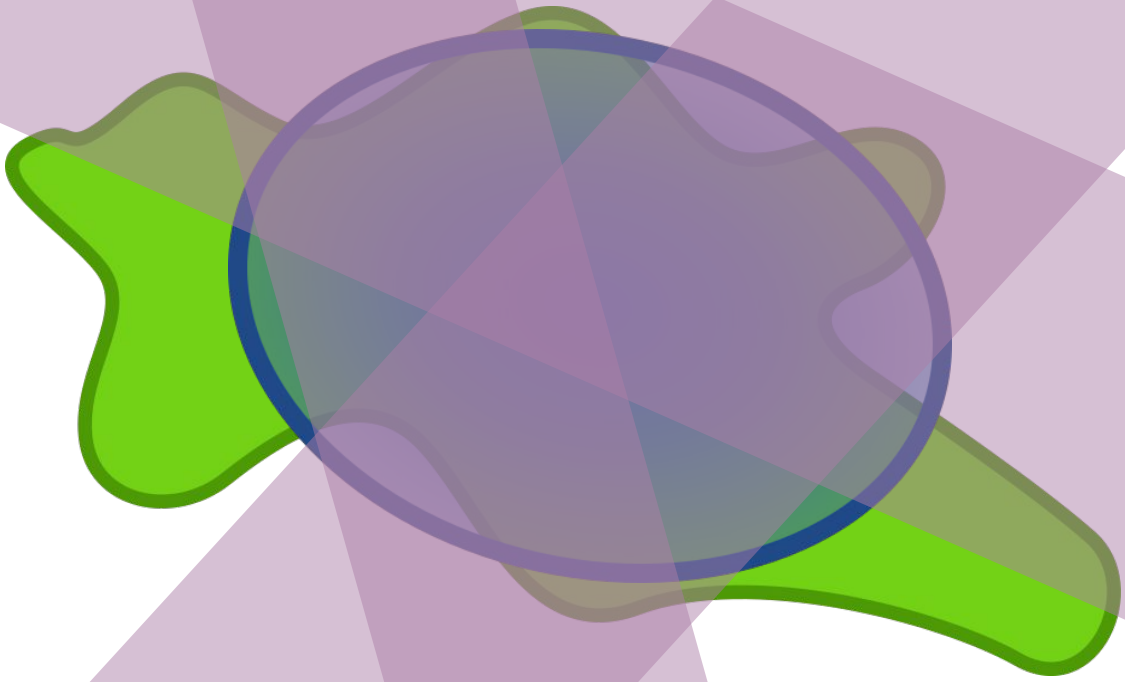
Localized Narrative



[Pont-Tuset et al. ECCV 2020]

Hints

- Image labels
- Image captions
- Video labels
- Transfer across images
- Transfer across classes
- Click inside object
- Size-from-center-click
- Scribbles
- Eye gaze
- Object bounding boxes
- Objects extreme points



Priors + Hints + Application

=

Weakly supervised paper

Checklist when (virtualy) visiting a poster:

- **Which priors** are being used ?
- How are these encoded ?

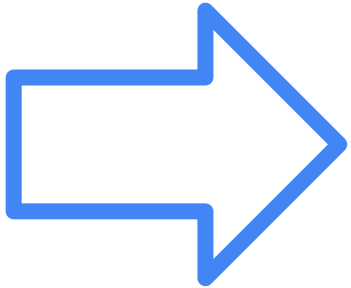
- **Which information source** is used ?
- Why was not that source exploited before ?

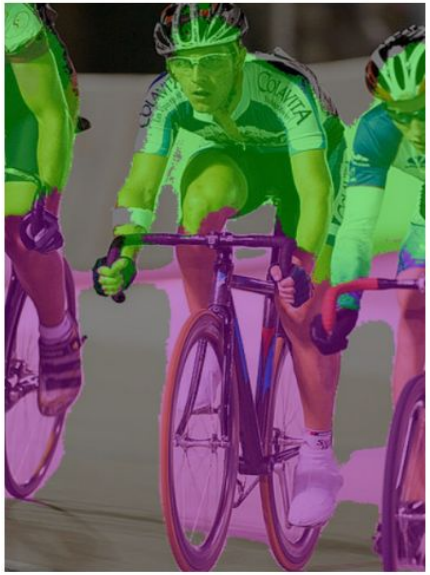
From image-level labels to pixel-level labels

(training to localize)

From image-level labels to pixel-level labels

Image-level labels
at training time





Pixel-level labels
at test time

Two main approaches:

- Gray box
  - [Simonyan et. al 2014]
  - [Smilkov et al. 2017]
  - [Fong & Vedaldi 2017]
- White box [Zhou et al. 2016]

From classifier to pixels, two main approaches:

- Gray box
  - **[Simonyan et. al 2014]**
  - [Smilkov et al. 2017]
  - [Fong & Vedaldi 2017]
- White box [Zhou et al. 2016]

Step 1: Train an image classifier

From classifier to pixels, two main approaches:

- Gray box
  - **[Simonyan et. al 2014]**
  - [Smilkov et al. 2017]
  - [Fong & Vedaldi 2017]
- White box [Zhou et al. 2016]

Step 1: Train an image classifier
Step 2: given an image with known label, compute

$$\left.\frac{\partial S_c}{\partial I}\right|_{I_0}$$

("how much does changes in the input image affect the score?")

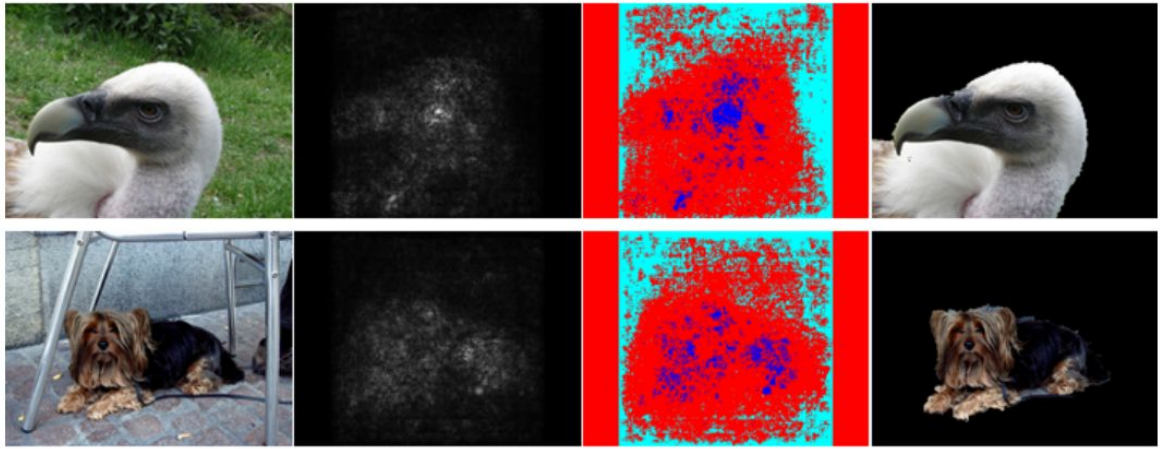
From classifier to pixels, two main approaches:

- Gray box
  - **[Simonyan et. al 2014]**
  - [Smilkov et al. 2017]
  - [Fong & Vedaldi 2017]
- White box [Zhou et al. 2016]

Step 1: Train an image classifier
Step 2: given an image with known label, compute

$$\left.\frac{\partial S_c}{\partial I}\right|_{I_0}$$

("how much does changes in the input image affect the score?")



(no quantitative experiments)

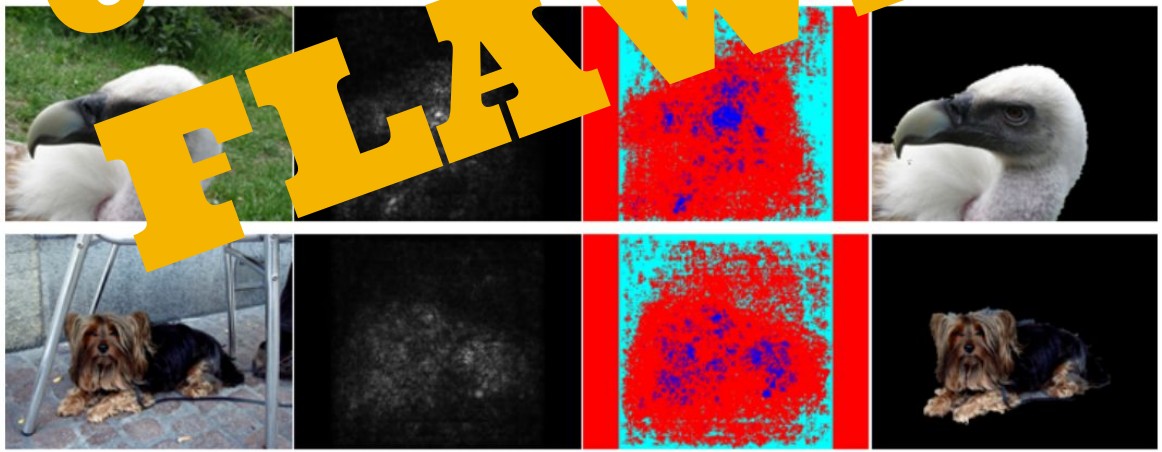
From classifier to pixels, two main approaches:

- Gray box
  - **[Simonyan et. al 2014]**
  - [Smilkov et al. 2017]
  - [Fong & Vedaldi 2017]
- White box [Zhou et al. 2016]

Step 1: Train an image classifier
Step 2: given an image with known label, compute

$$\frac{\partial S_c}{\partial I}\Big|_{I_0}$$

("how much do changes in the input image affect the score")



Conceptually FLAWED

(does not work for linear functions,
does not work for good classifiers)

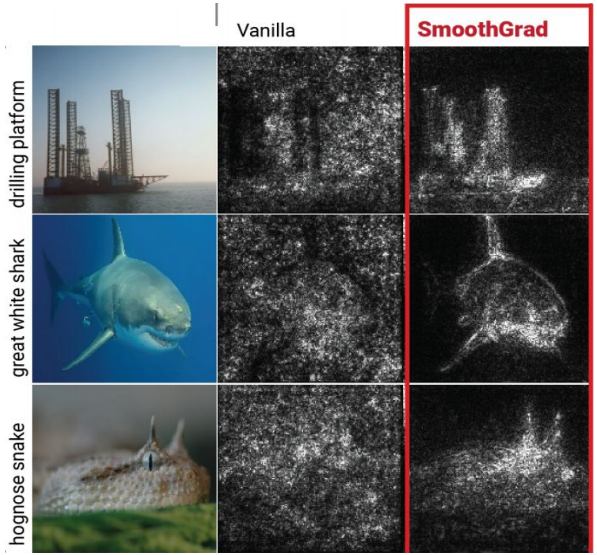
From classifier to pixels, two main approaches:

- Gray box
  - ~~[Simonyan et. al 2014]~~
  - **[Smilkov et al. 2017]**
  - [Fong & Vedaldi 2017]
- White box [Zhou et al. 2016]

Step 1: Train an image classifier
Step 2: **Add small random noise** to image
Step 3: given known label,
compute $\left.\frac{\partial S_c}{\partial I}\right|_{I_0}$

("how much does changes in the input image affect the score?")

Step 4: **Repeat** 50 times step 2&3, **average result**



(no quantitative experiments either)

From classifier to pixels, two main approaches:

- Gray box
  - ~~[Simonyan et. al 2014]~~
  - ~~[Smilkov et al. 2017]~~
  - ~~[Fong & Vedaldi 2017]~~
- White box [Zhou et al. 2016]

Step 1: train an image classifier
Step 2: given image with known label
Step 3: **pray for the best**

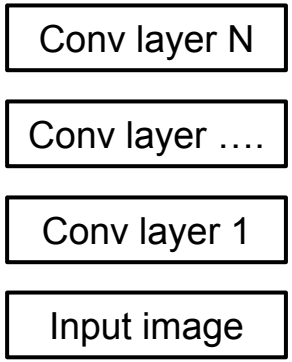
Key issue: Asking apples from the oranges tree

Classifier was not built nor trained to segment objects

From classifier to pixels, two main approaches:

- ~~Gray box~~
  - ~~[Simonyan et. al 2014]~~
  - ~~[Smilkov et al. 2017]~~
  - ~~[Fong & Vedaldi 2017]~~
- **White box** [Zhou et al. 2016]

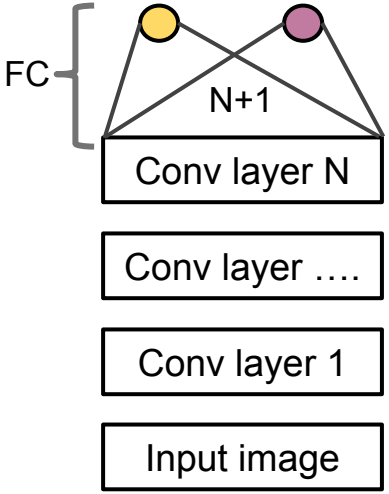
Step 1: change the network architecture
Step 2: train an image classifier
Step 3: apply over new image, extract the segment



From classifier to pixels, two main approaches:

- ~~Gray box~~
  - ~~[Simonyan et. al 2014]~~
  - ~~[Smilkov et al. 2017]~~
  - ~~[Fong & Vedaldi 2017]~~
- **White box** [Zhou et al. 2016]

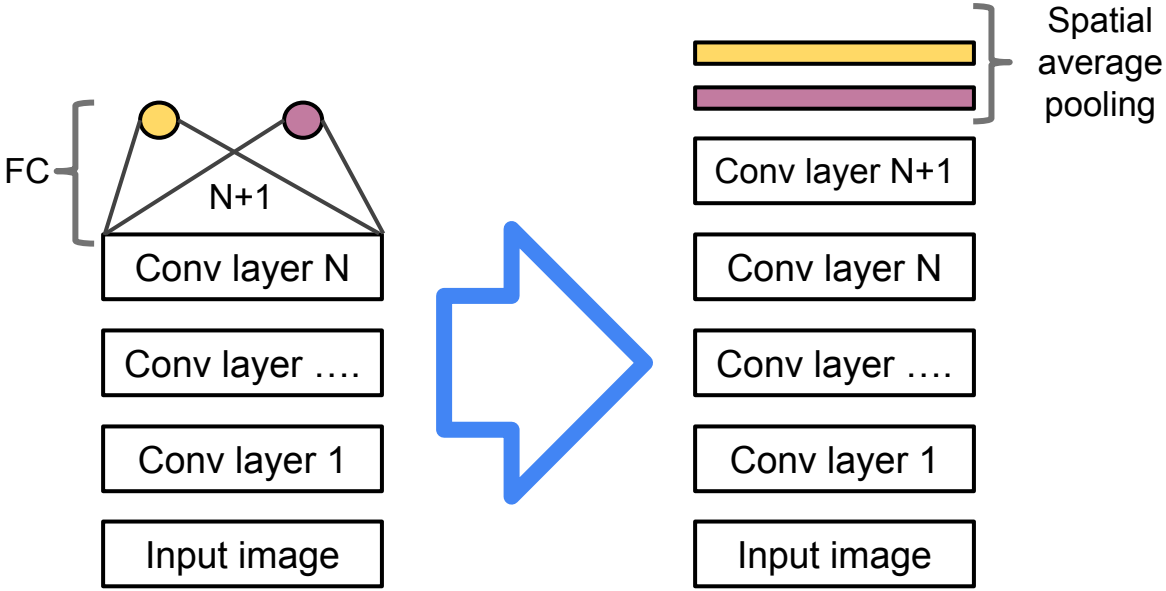
Step 1: change the network architecture
Step 2: train an image classifier
Step 3: apply over new image, extract the segment



From classifier to pixels, two main approaches:

- ~~Gray box~~
  - ~~[Simonyan et. al 2014]~~
  - ~~[Smilkov et al. 2017]~~
  - ~~[Fong & Vedaldi 2017]~~
- **White box** [Zhou et al. 2016]

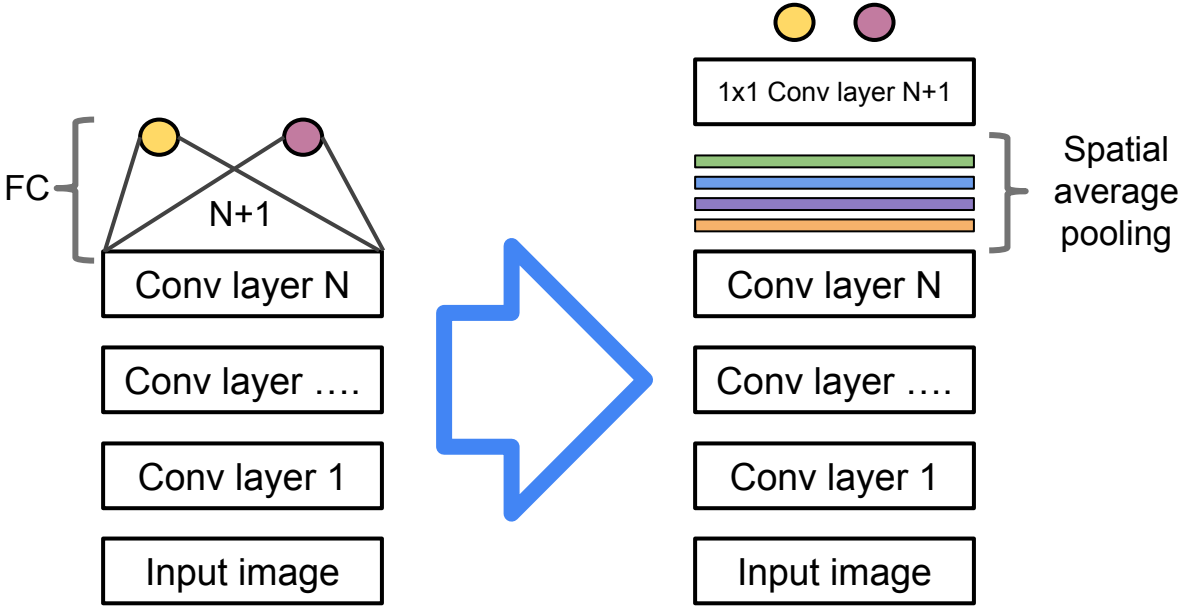
Step 1: **change** the network architecture
Step 2: train an image classifier
Step 3: apply over new image, extract the segment

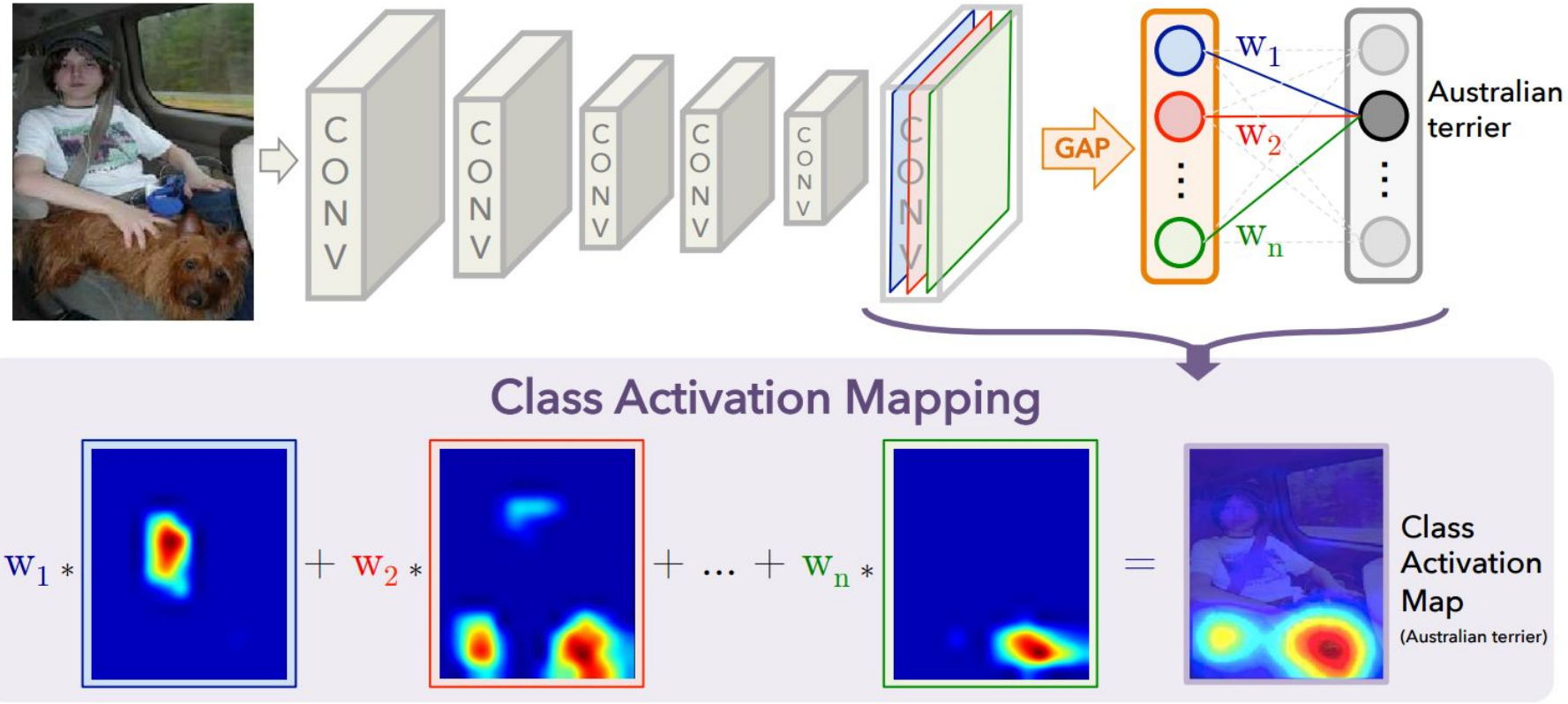


From classifier to pixels, two main approaches:

- ~~Gray box~~
  - ~~[Simonyan et. al 2014]~~
  - ~~[Smilkov et al. 2017]~~
  - ~~[Fong & Vedaldi 2017]~~
- **White box** [Zhou et al. 2016]

Step 1: change the network architecture
Step 2: train an image classifier
Step 3: apply over new image, extract the segment





(Approach is known as GAP or CAM)

[Zhou et al. 2016]

We are now training a classifier architecture designed to generate per-class maps

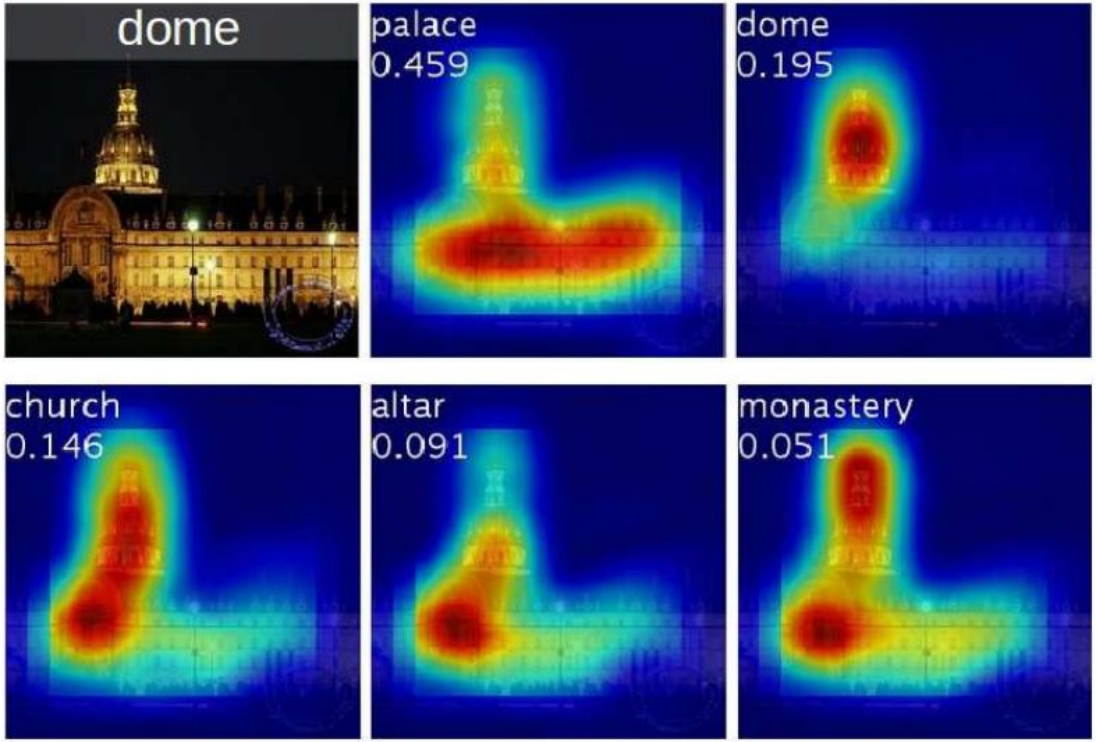
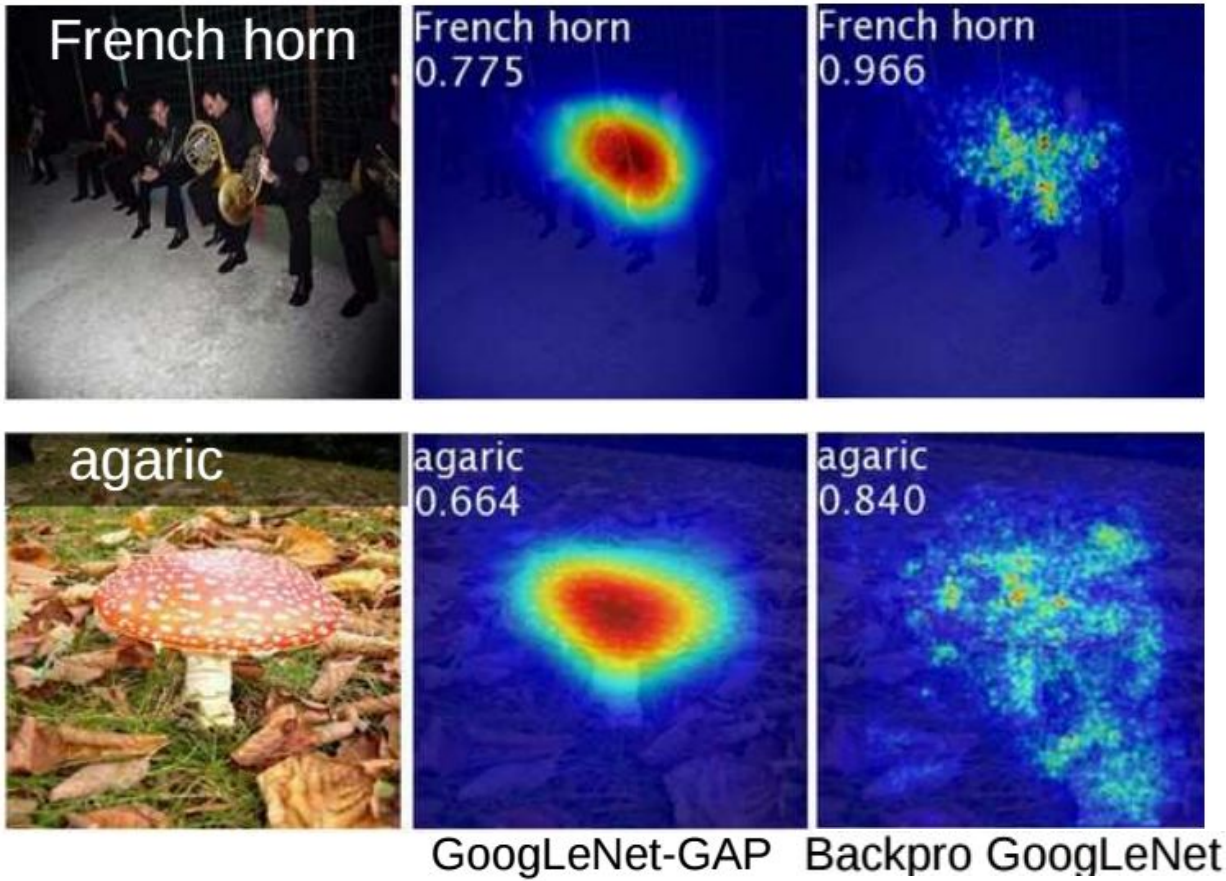
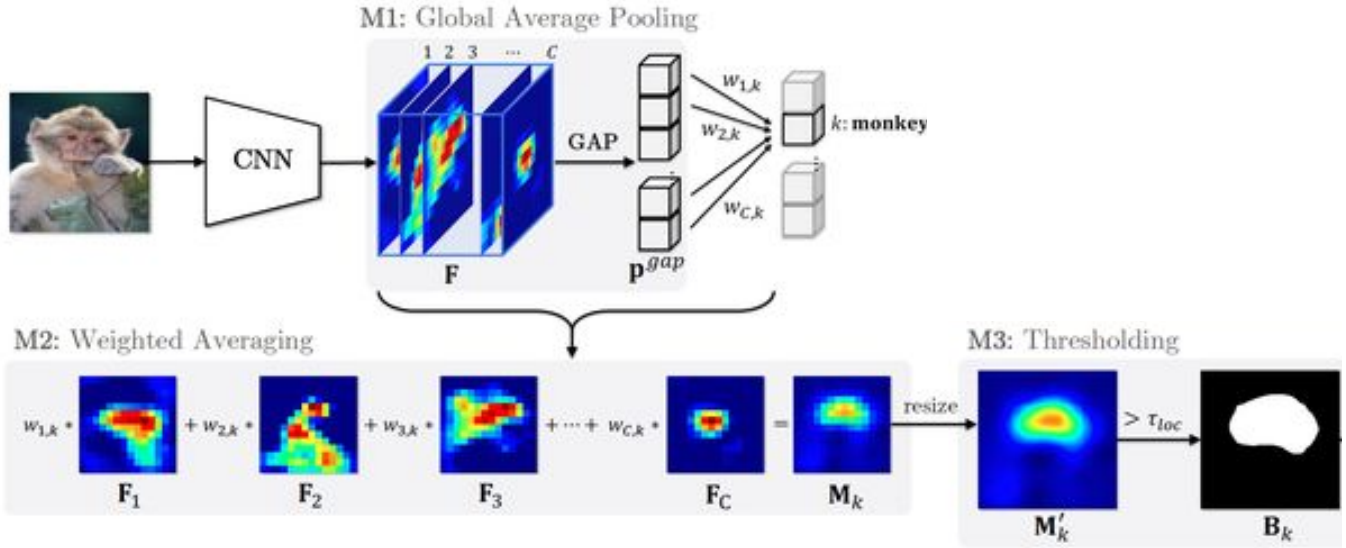


Figure 4. Examples of the CAMs generated from the top 5 predicted categories for the given image with ground-truth as dome.

We are now training a classifier architecture designed to generate per-class maps



The class maps are thresholded, to generate training data

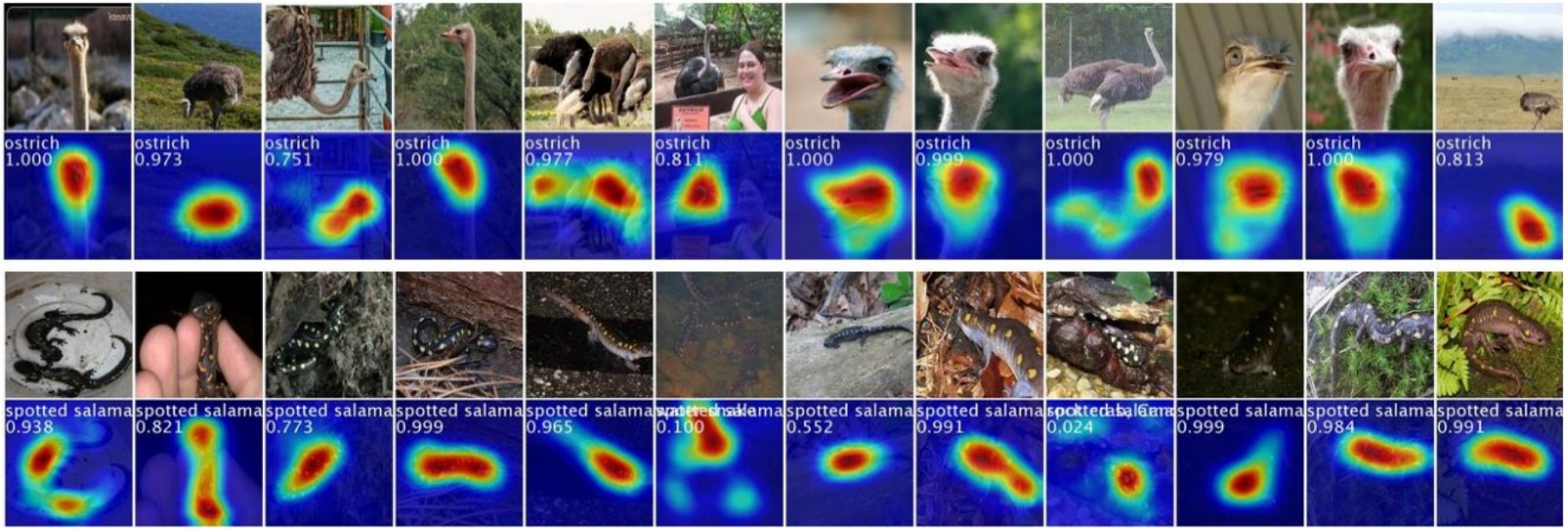


High value scores → Foreground areas
Low value scores → Background areas

(This confident regions are commonly named “seeds”)

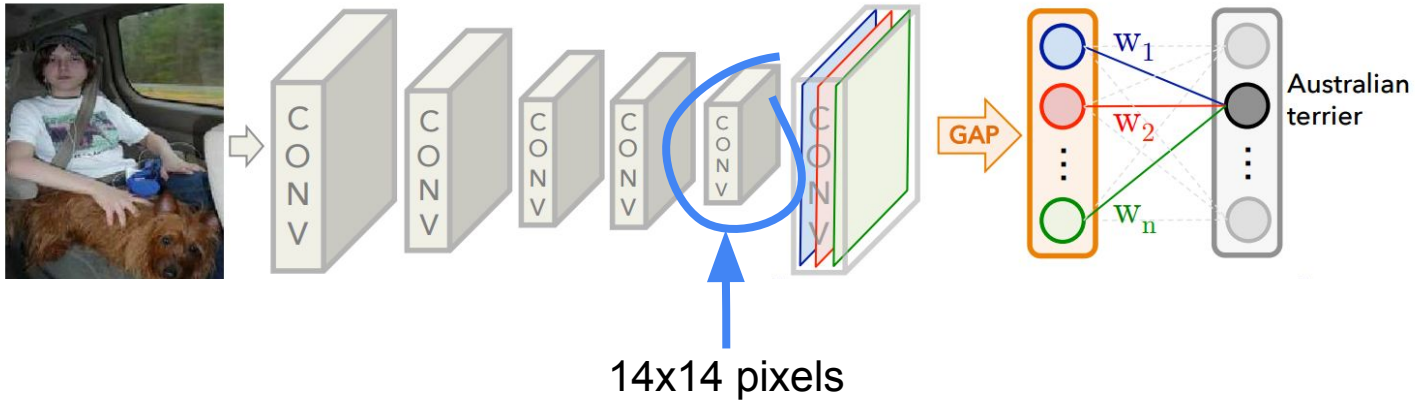
Good, but not good enough:

1. Resulting masks are not sharp
2. Focused on discriminative area only



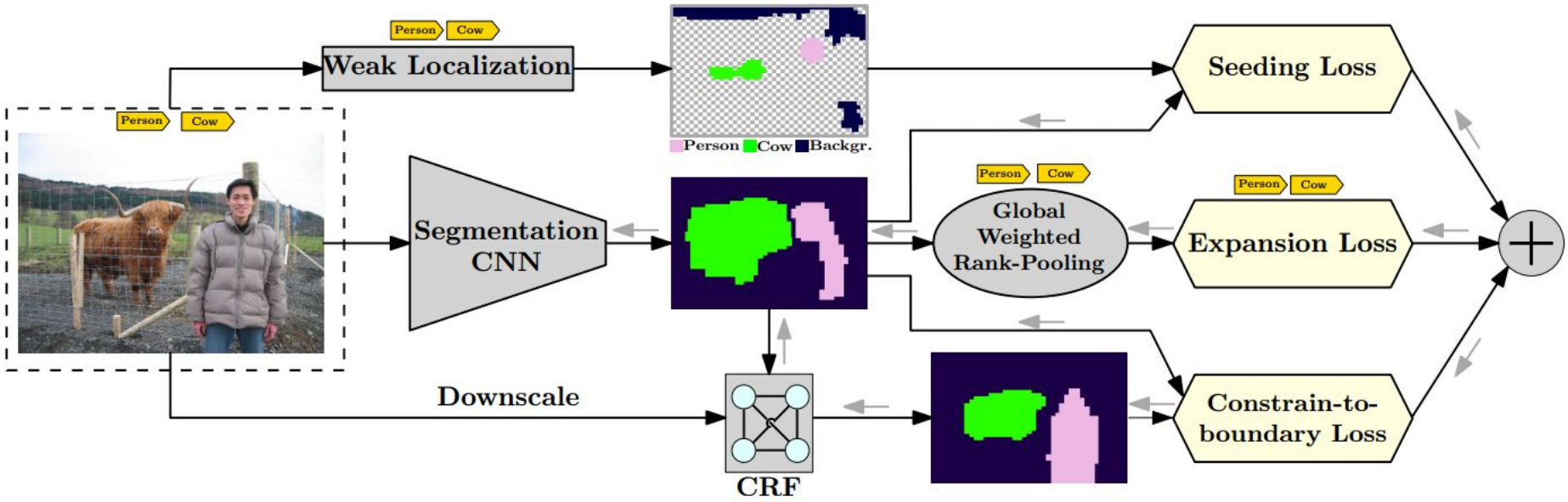
Good, but not good enough:

1. Resulting masks are not sharp
   ⇢ Increase top layers resolution
2. Focused on discriminative area only



Good, but not good enough:

1. Resulting masks are not sharp
   → **Use boundary-aware propagation**
2. Focused on discriminative area only



GAP^{++} + Seeds + **CRF loss**

[Kolesnikov & Lampert ECCV 2016]

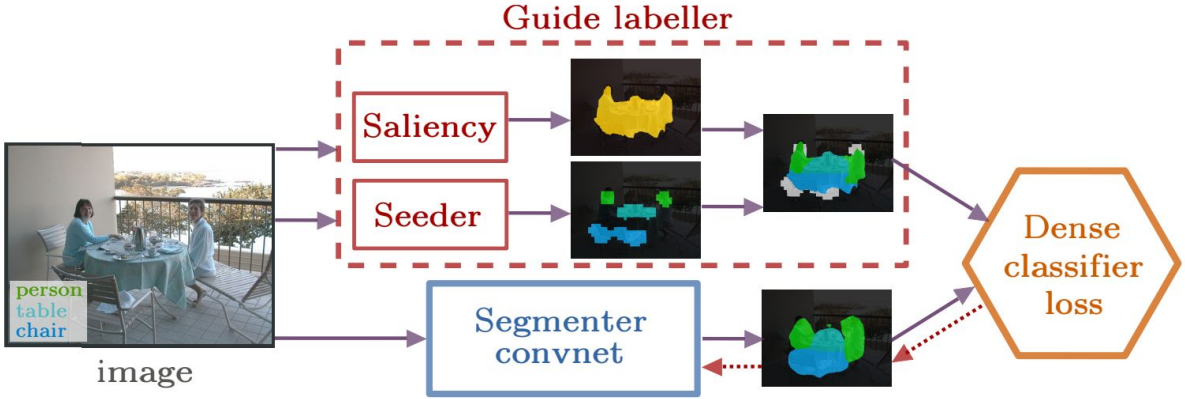
Good, but not good enough:

1. Resulting masks are not sharp
   → **Use transfer learning**
2. Focused on discriminative area only

"We could also use saliency cues"

image labels

saliency

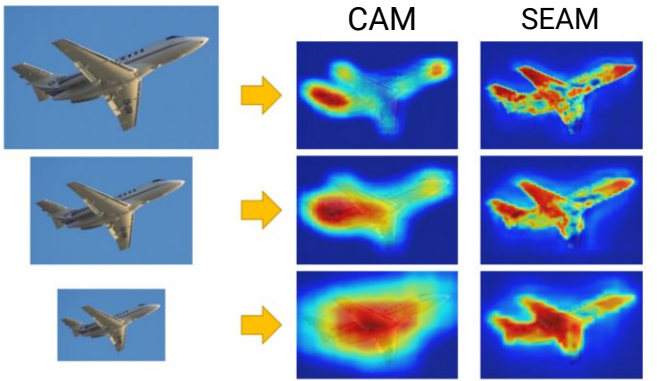


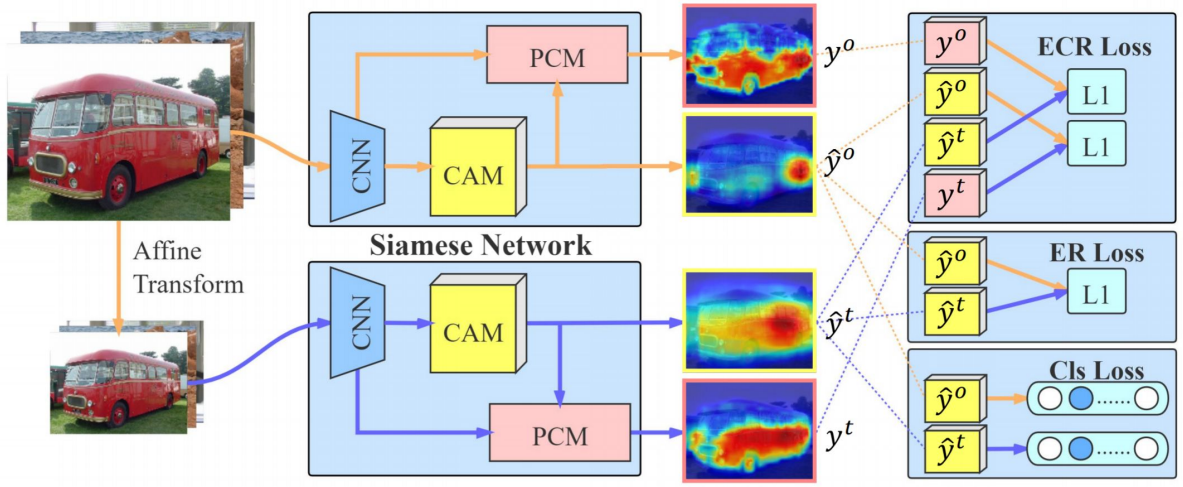
[Oh et al. CVPR 2017]

Good, but not good enough:

1. Resulting masks are not sharp
   → **Enforce equivariance**
2. Focused on discriminative area only

"Activation maps should be equivariant to the input"





[Wang et al. CVPR 2020]

Good, but not good enough:

1. Resulting masks are not sharp
2. Focused on discriminative area only

   ⇢ **Let us play peek-a-boo!**

Good, but not good enough:

1. Resulting masks are not sharp
2. Focused on discriminative area only

   → **Let us play peek-a-boo!**

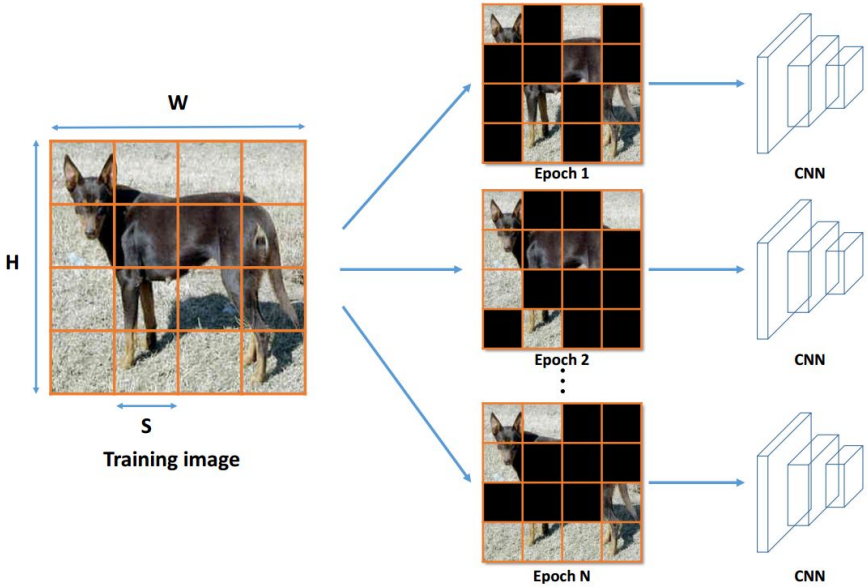


Figure 2. **Approach overview.** For each training image, we divide it into a grid of $S \times S$ patches. Each patch is then randomly hidden with probability p_{hide} and given as input to a CNN to learn image classification. The hidden patches change randomly across different epochs, which forces the network to focus on different parts of the object for learning image classification.

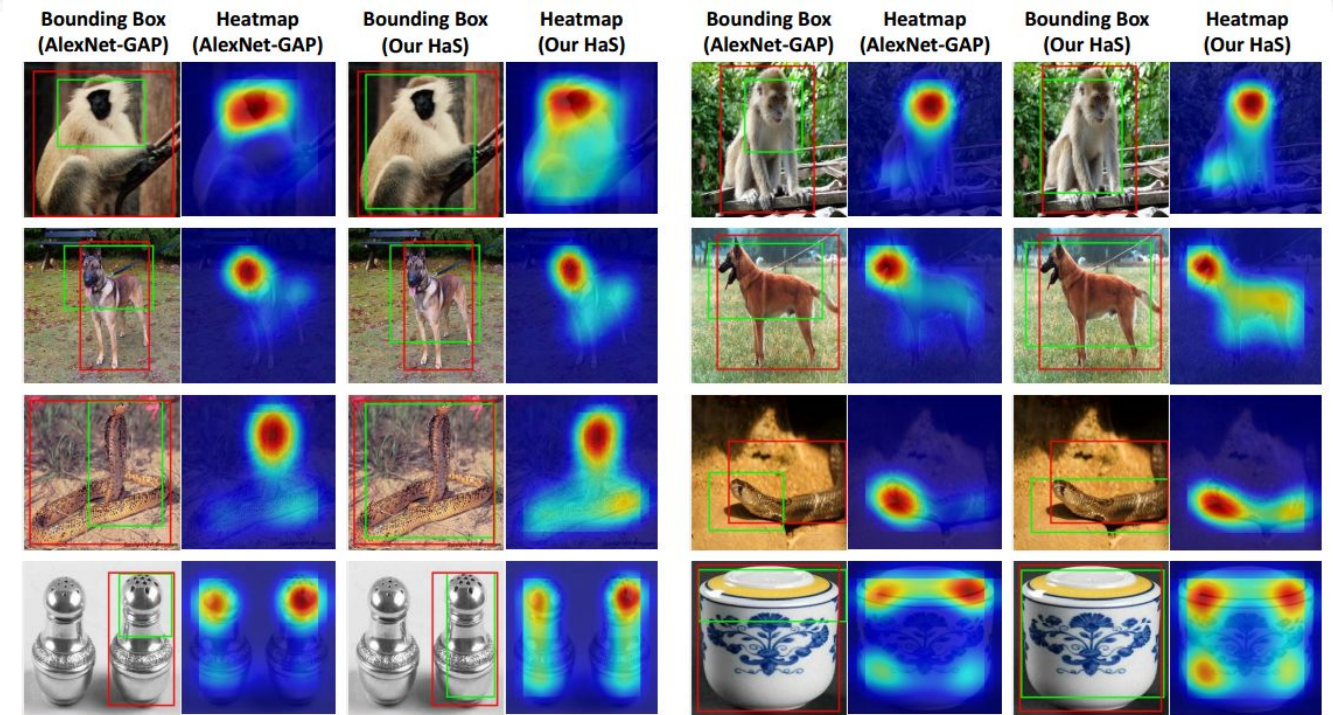
(identical to SpatialDropout [Tompson et al. 2015])

[Singh & Lee 2017]

Good, but not good enough:

1. Resulting masks are not sharp
2. Focused on discriminative area only

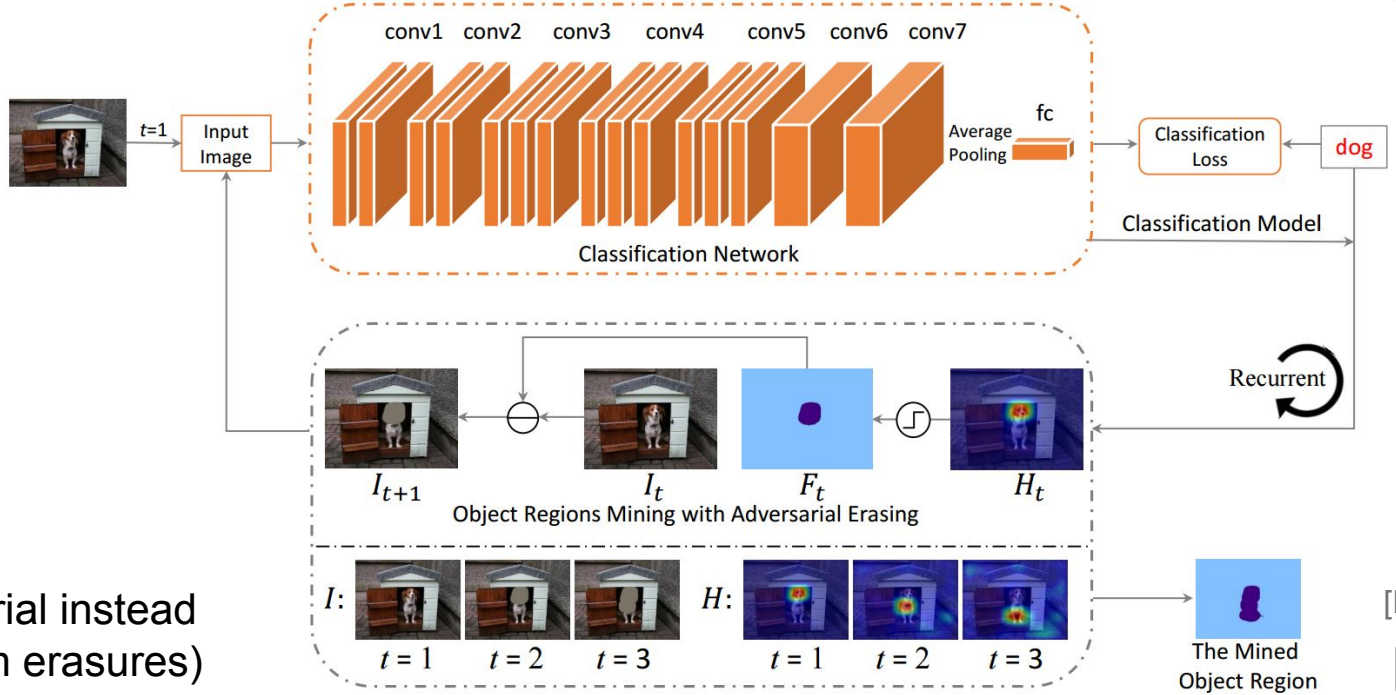
   ⇢ **Let us play peek-a-boo!**



Good, but not good enough:

1. Resulting masks are not sharp
2. Focused on discriminative area only

   → **Let us play <u>adversarial</u> peek-a-boo!**



(adversarial instead of random erasures)

(Similar in spirit in to [Fong & Vedaldi 2017])

[Wei et al. 2017]

Good, but not good enough:

1. Resulting masks are not sharp
2. Focused on discriminative area only

   → **Let us play adversarial peek-a-boo!**

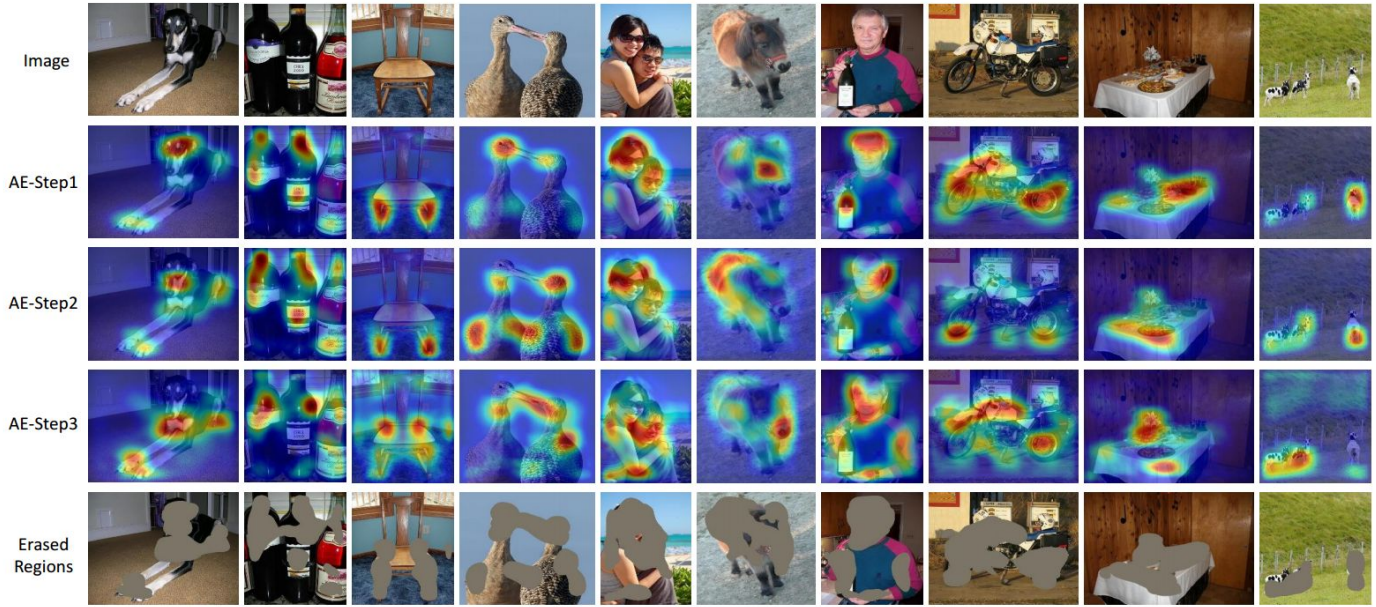


Figure 6. Examples of mined object regions produced by the proposed adversarial erasing approach. The second to fourth rows show the produced heatmaps, where the discriminative regions are highlighted. The images with erased regions are shown in the last row in gray.

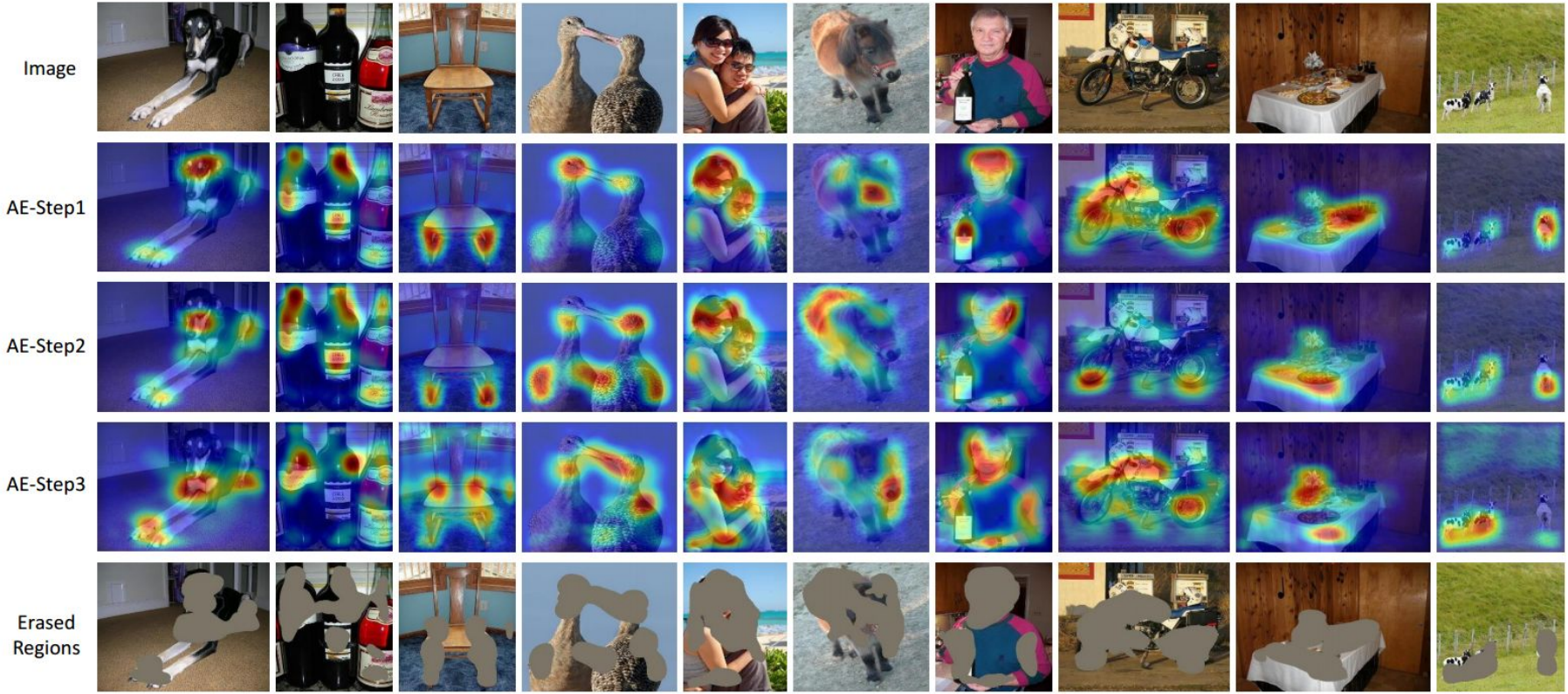


Figure 6. Examples of mined object regions produced by the proposed adversarial erasing approach. The second to fourth rows show the produced heatmaps, where the discriminative regions are highlighted. The images with erased regions are shown in the last row in gray.

[Wei et al. 2017]

Good, but not good enough:

1. Resulting masks are not sharp
2. Focused on discriminative area only

   → **Let us play adversarial peek-a-boo!**

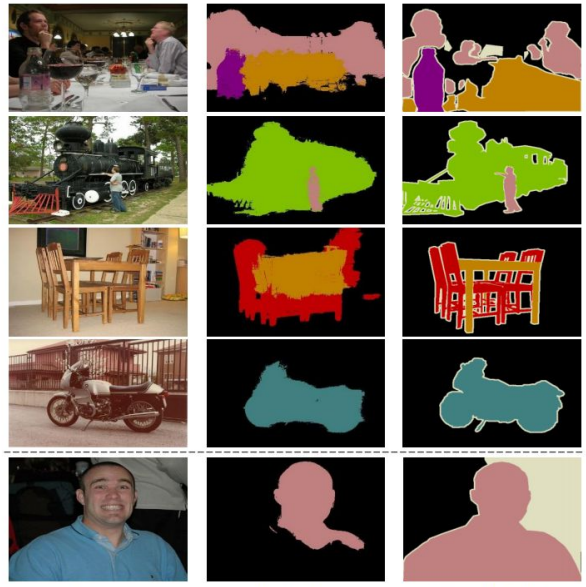


Figure 4. Qualitative segmentation results on the VOC 2012 *val* set. One failure case is shown in the last row.

Table 2. Comparison of weakly-supervised semantic segmentation methods on VOC 2012 *test* set.

Methods	Training Set	mIoU
Supervision: Box		
WSSL (ICCV 2015) [16]	10K	62.2
BoxSup (ICCV 2015) [3]	10K	64.2
Supervision: Image-level Labels (* indicates methods implicitly use pixel-level supervision)		
MIL-seg* (CVPR 2015) [19]	700K	40.6
SN_B* (PR 2016) [28]	10K	43.2
TransferNet* (CVPR 2016) [7]	70K	51.2
AF-MCG* (ECCV 2016) [20]	10K	55.5
Supervision: Image-level Labels		
MIL-FCN (ICLR 2015) [18]	10K	24.9
CCNN (ICCV 2015) [17]	10K	35.6
MIL-sppxl (CVPR 2015) [19]	700K	35.8
MIL-bb (CVPR 2015) [19]	700K	37.0
EM-Adapt (ICCV 2015) [16]	10K	39.6
DCSM (ECCV 2016) [24]	10K	45.1
BFBP (ECCV 2016) [23]	10K	48.0
STC (PAMI 2016) [29]	50K	51.2
SEC (ECCV 2016) [10]	10K	51.7
AF-SS (ECCV 2016) [20]	10K	52.7
Supervision: Image-level Labels		
AE-PSL (ours)	10K	**55.7**

Takeaways:

- Training a classifier, and hoping to get segmentations out of it ✗
- Using spatial continuity priors will help, but to a limit ~
- The training can be modified to promote covering the object extent ✓
- With careful changes it is possible to obtain significant improvements ~

⇒ **How far can we go from image-level to pixel-level labels ?**

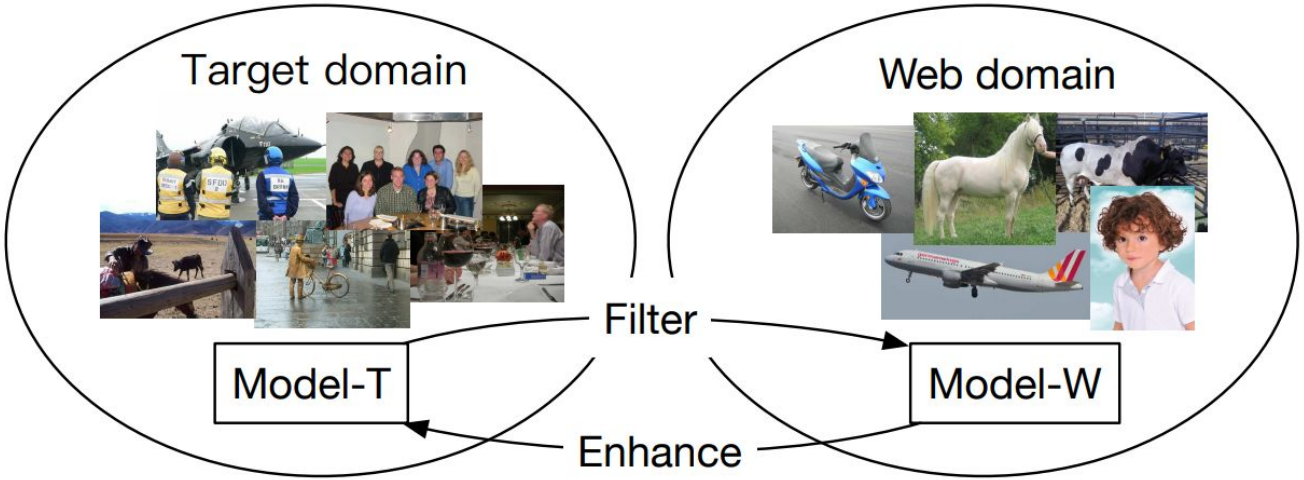
From image-level labels to pixel-level labels

(webly supervised)

Webly Supervised Learning

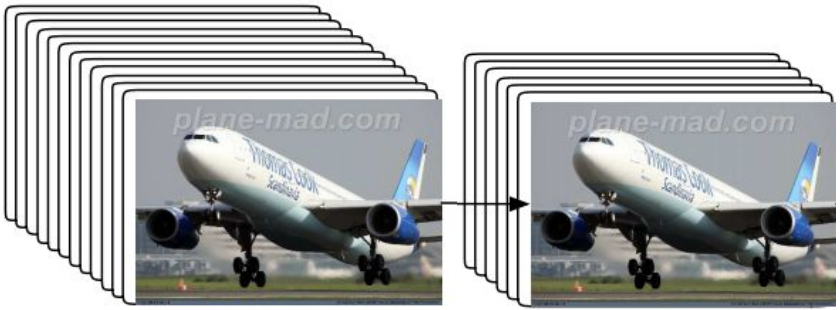
[Shen CVPR 18, Jin CVPR 17, Chen ICCV 15]

1. Train image classifier Model-T



[Shen CVPR 18]

1. Train image classifier Model-T
2. Filter web images with Model-T

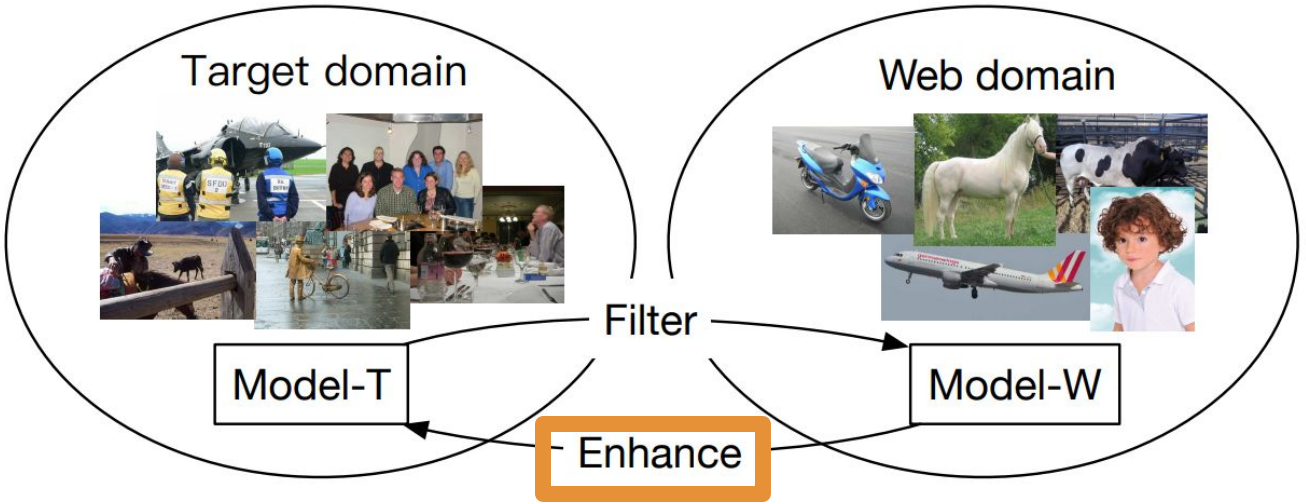


[Shen CVPR 18]

1. Train image classifier Model-T
2. Filter web images with Model-T
3. Train pixel labeler Web-SEC using web images

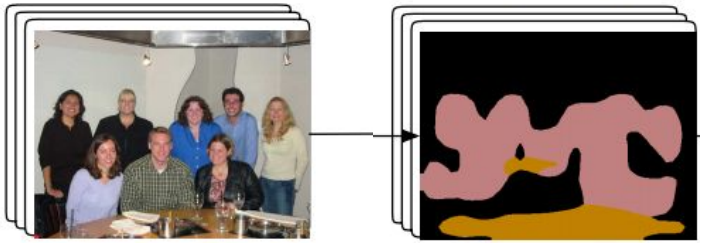
[Shen CVPR 18]

1. Train image classifier Model-T
2. Filter web images with Model-T
3. Train pixel labeler Web-SEC using web images
4. Apply Web-SEC over the target images



[Shen CVPR 18]

1. Train image classifier Model-T
2. Filter web images with Model-T
3. Train pixel labeler Web-SEC using web images
4. Apply Web-SEC over the target images
5. Use Model-T to obtain a second labeling over target images

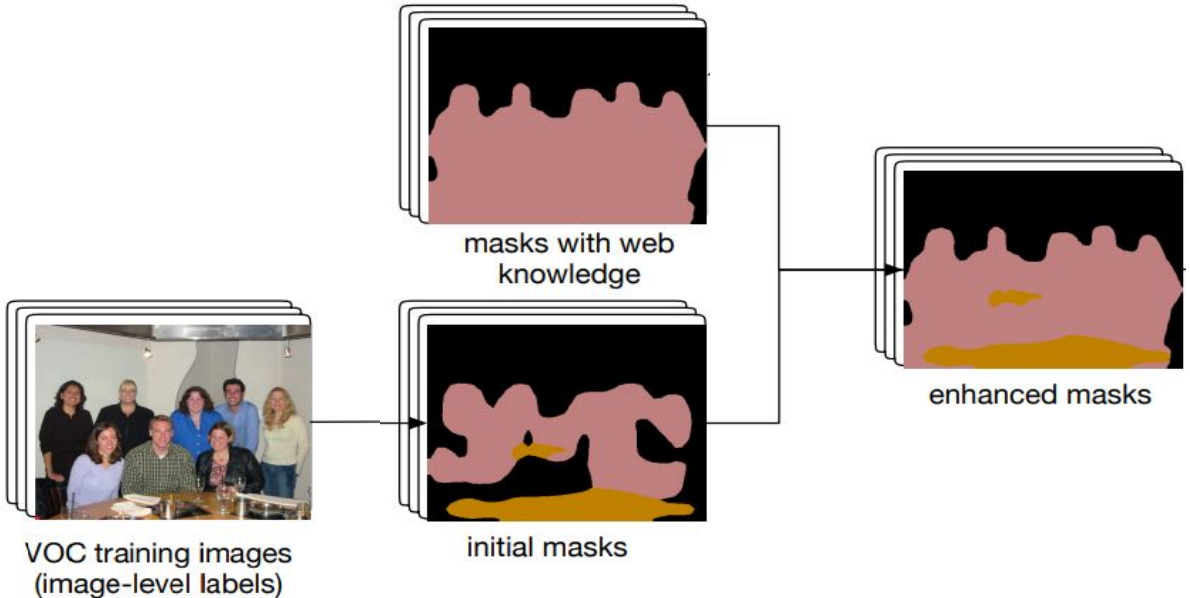


VOC training images
(image-level labels)

initial masks

[Shen CVPR 18]

1. Train image classifier Model-T
2. Filter web images with Model-T
3. Train pixel labeler Web-SEC using web images
4. Apply Web-SEC over the target images
5. Use Model-T to obtain a second labeling over target images
6. Merge (4) and (5), train final model



[Shen CVPR 18]

Webly supervised learning can be quite effective

Method	val	test	Extra Supervision
Chen *et al.* [2]	**63.7**	**66.4**	Fully supervised
Lin *et al.* [16]	63.1	-	Scribble
Dai *et al.* [6]	62.0	64.6	Bounding box+MCG
Oh *et al.* [21]	55.7	56.7	Bounding box
Bearman *et al.* [1]	46.1	-	Point
Wei *et al.* [29]	55.0	55.7	Supervised saliency
STC [30]	49.8	51.2	Supervised saliency
EM-Adapt [22]	33.8	39.6	-
CCNN [23]	35.3	35.6	-
SEC [13]	50.7	51.7	-
Hong *et al.* [10]	58.1	58.7	-
Ours-VGG16	58.8	60.2	Web
Ours-Res50	**63.0**	**63.9**	Web

Table 1: Comparison with methods using other supervisions.

input prediction

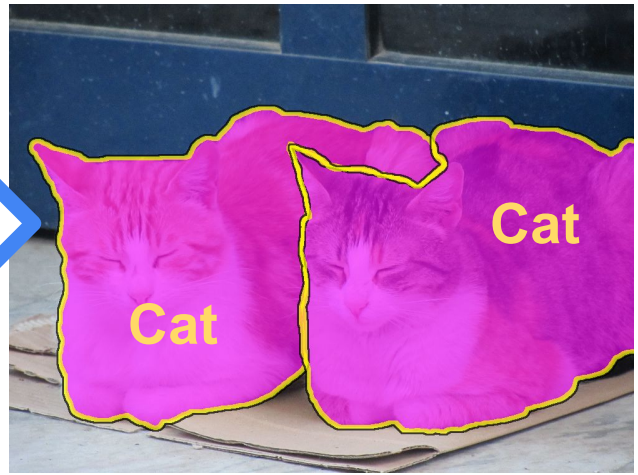
[Shen CVPR 18]

From image-level labels to pixel-level labels

(instance segmentation)

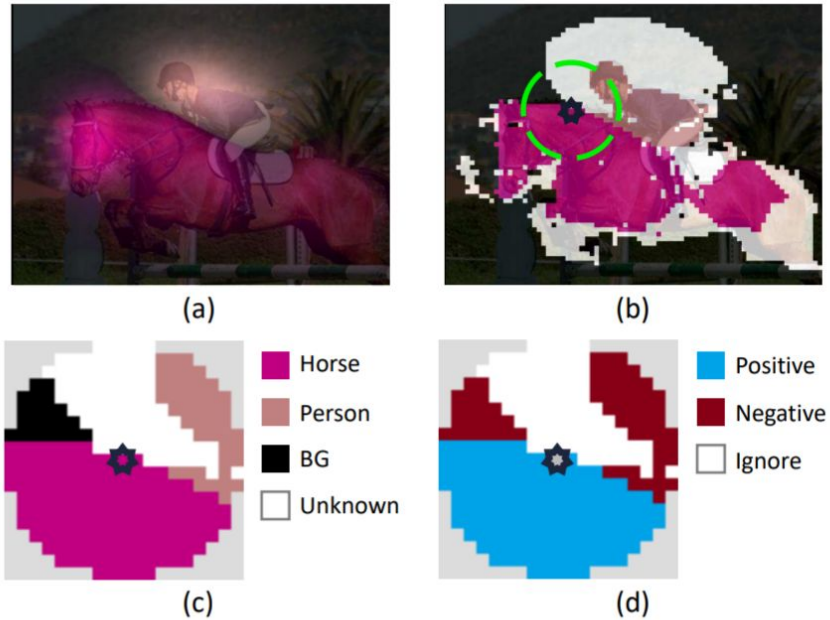
From image-level labels to pixel-level instance labels

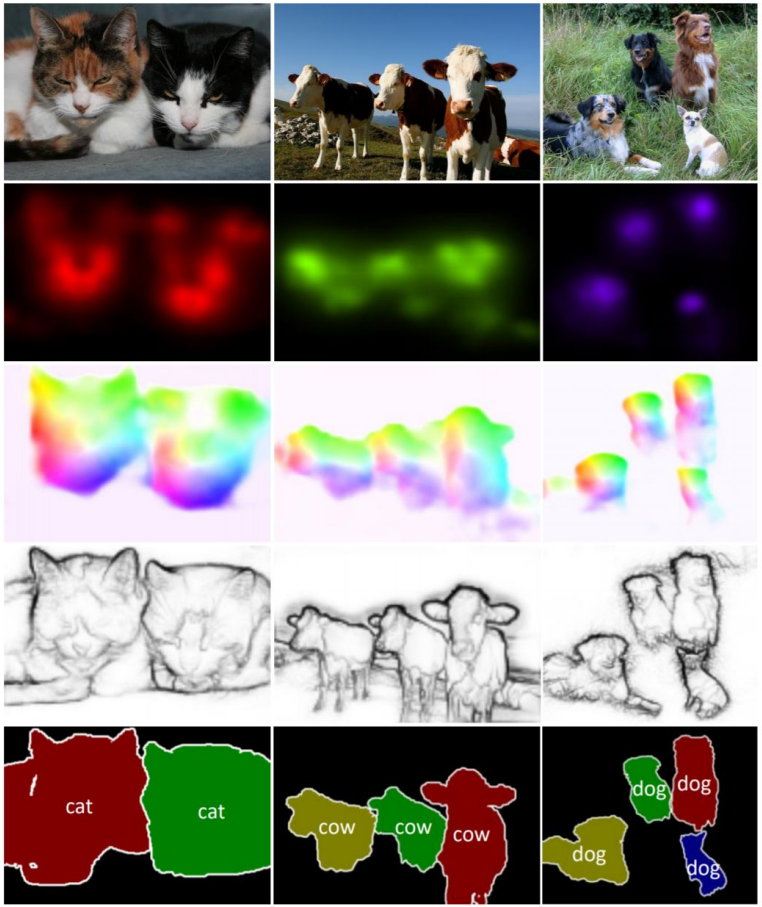
Image-level labels
at training time



Pixel-level per
instance labels
at test time

CAM □ Boundaries detector
CAM □ Centroid estimator
CAM + Boundaries + centroids □ Instances masks

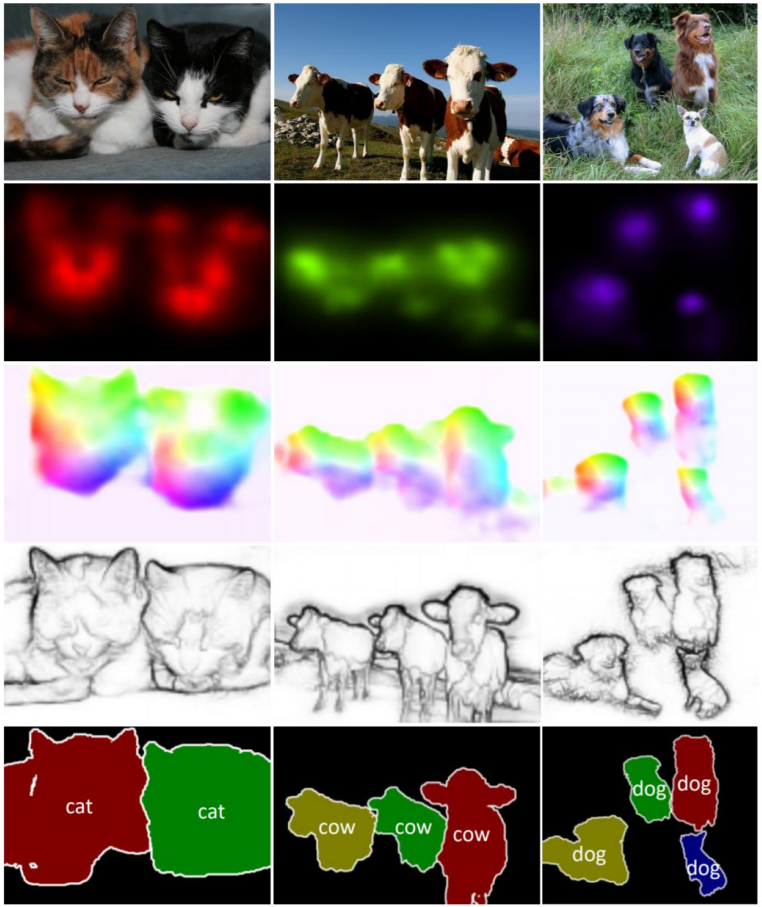




Only paper to go from *only* priors + image labels to instance segmentation.
No transfer from pre-trained saliency nor boundary detector.

[Ahn et al. CVPR 19]

CAM □ Boundaries detector
CAM □ Centroid estimator
CAM + Boundaries + centroids □ Instances masks



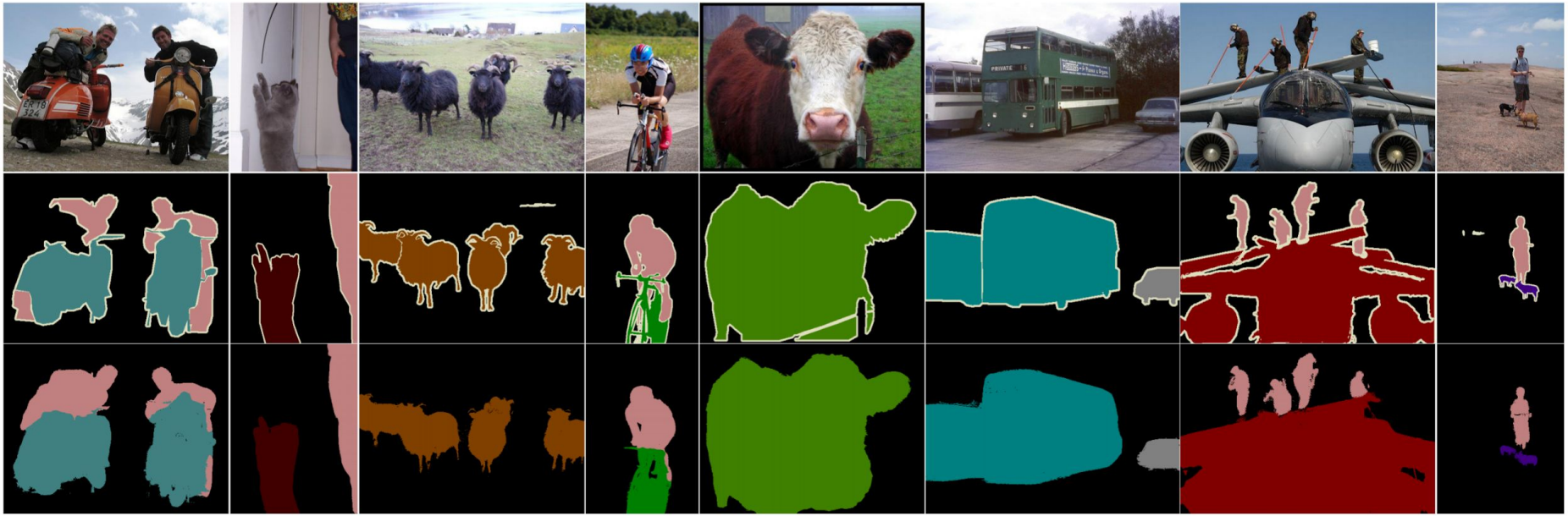
Method	Sup.	Extra data / Information	AP^r_{50}	AP^r_{70}
PRM [50]	$\mathcal{I}$	MCG [2]	26.8	-
SDI [22]	$\mathcal{B}$	BSDS [33]	44.8	-
SDS [16]	$\mathcal{F}$	MCG [2]	43.8	21.3
MRCNN [17]	$\mathcal{F}$	MS-COCO [29]	69.0	-
Ours-ResNet50	$\mathcal{I}$	-	46.7	23.5

Table 3. Instance segmentation performance on the PASCAL VOC 2012 *val* set. The supervision types (Sup.) indicate: $\mathcal{I}$–image-level label, $\mathcal{B}$–bounding box, and $\mathcal{F}$–segmentation label.

Only paper to go from *only* priors + image labels to instance segmentation.
No transfer from pre-trained saliency nor boundary detector.

[Ahn et al. CVPR 19]

CAM □ Boundaries detector
CAM □ Centroid estimator
CAM + Boundaries + centroids □ Instances masks



Only paper to go from *only* priors + image labels to instance segmentation.
No transfer from pre-trained saliency nor boundary detector.

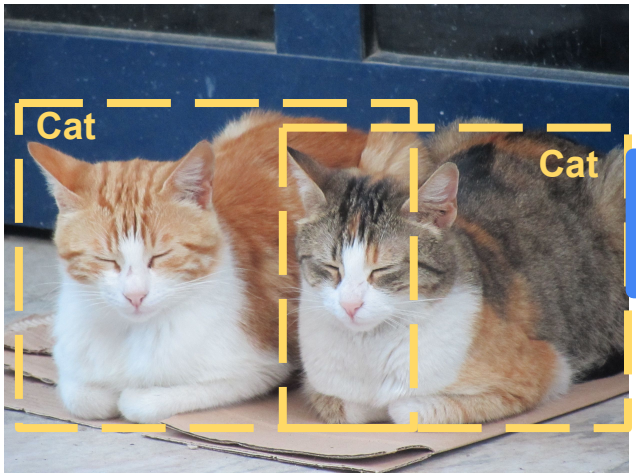
[Ahn et al. CVPR 19]

From box-level labels to pixel-level labels

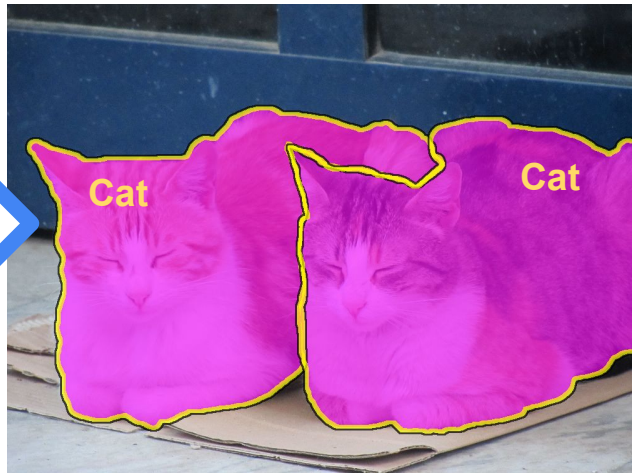
(boxly supervised)

(With slides kindly provided by *Tal Remez*)

From box-level labels to pixel-level instance labels



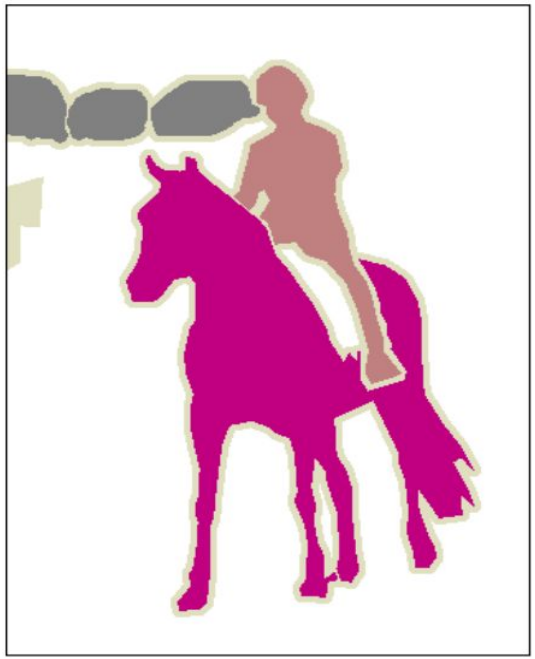
Box-level labels
at training time

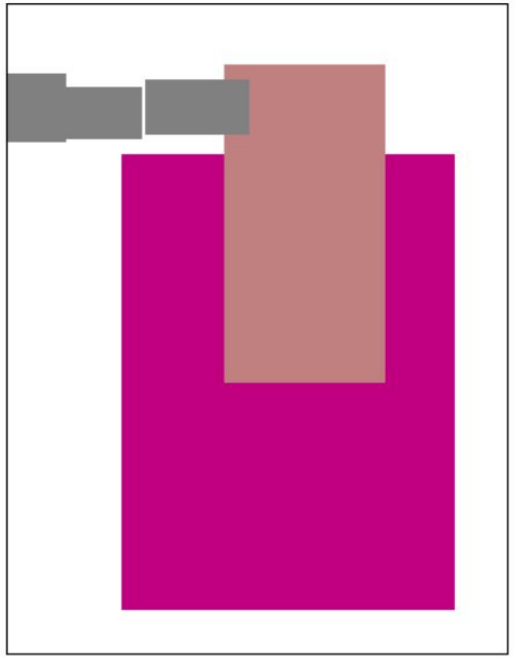


Pixel-level per
instance labels
at test time

What we want:

What we have:

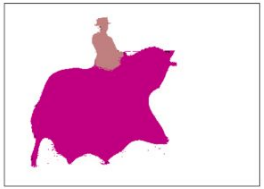




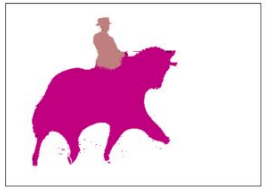
[Khoreva CVPR 17]

Convnets are robust to noise. Let us use that !

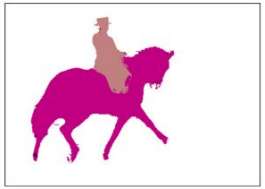
Example
input rectangles



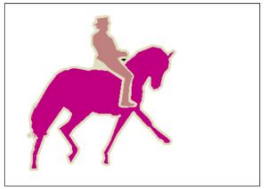
Output after
1 training round



After
5 rounds



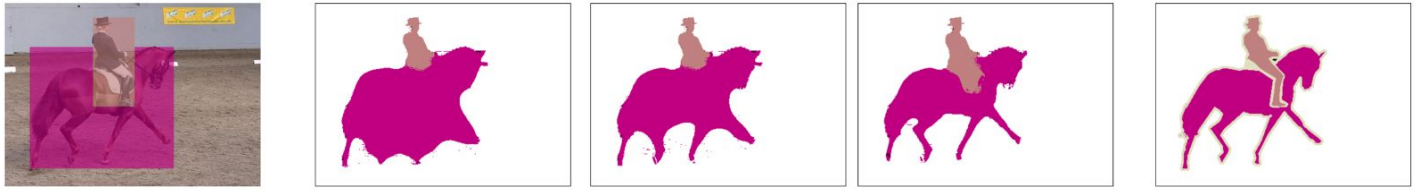
After
10 rounds



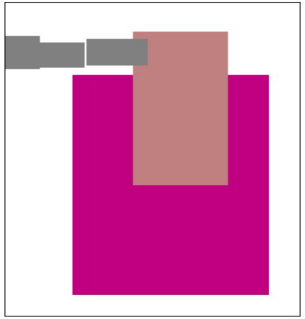
Ground
truth

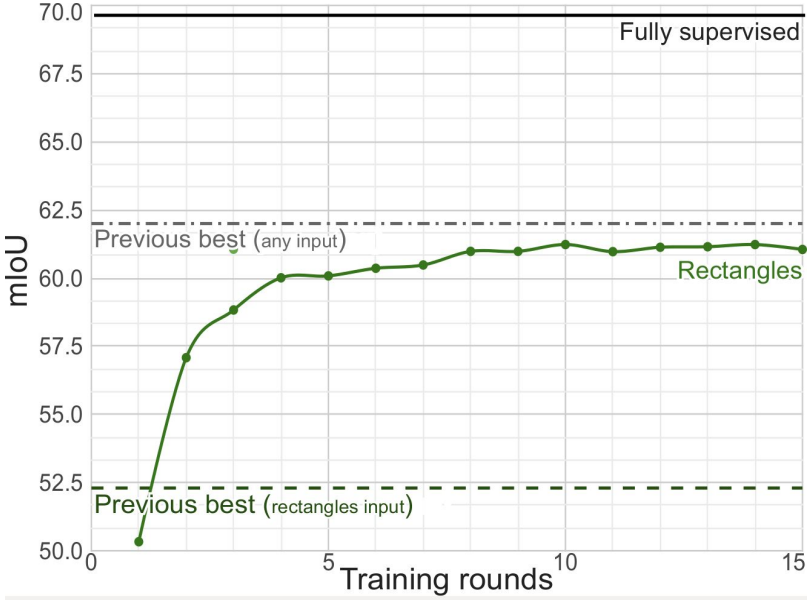
[Khoreva CVPR 17]

Convnets are robust to noise. Let us use that !



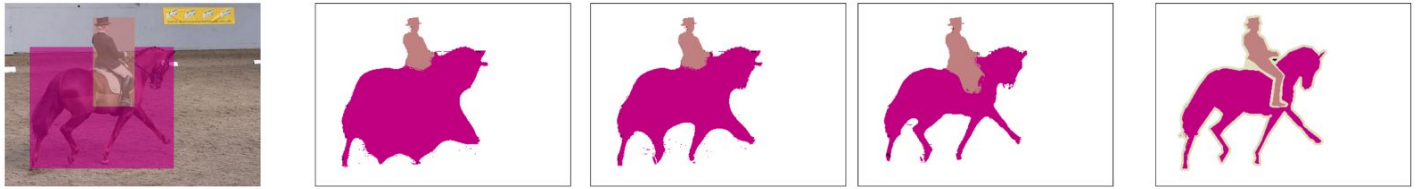
Example input rectangles | Output after 1 training round | After 5 rounds | After 10 rounds | Ground truth



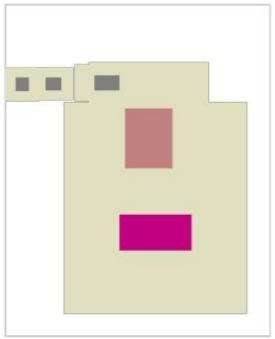


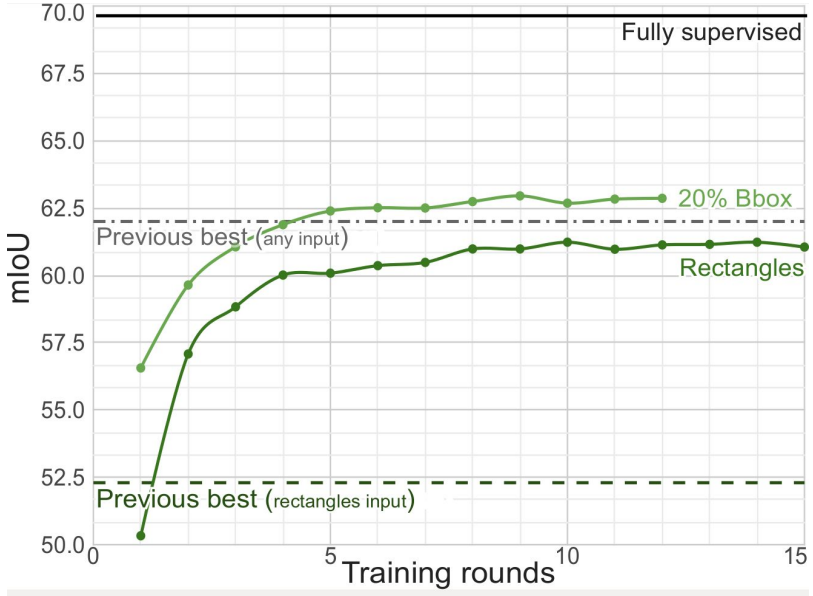
[Khoreva CVPR 17]

Convnets are robust to noise. Let us use that !



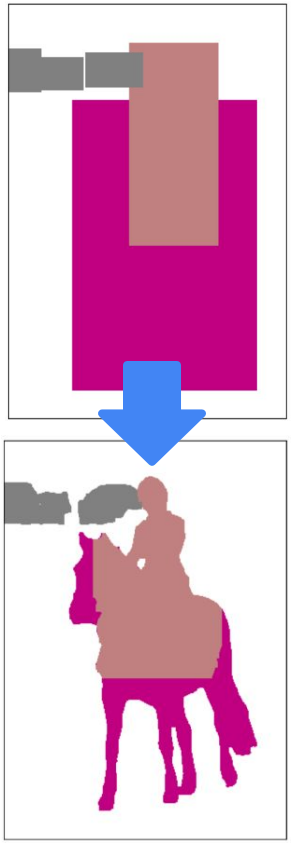
Example input rectangles | Output after 1 training round | After 5 rounds | After 10 rounds | Ground truth

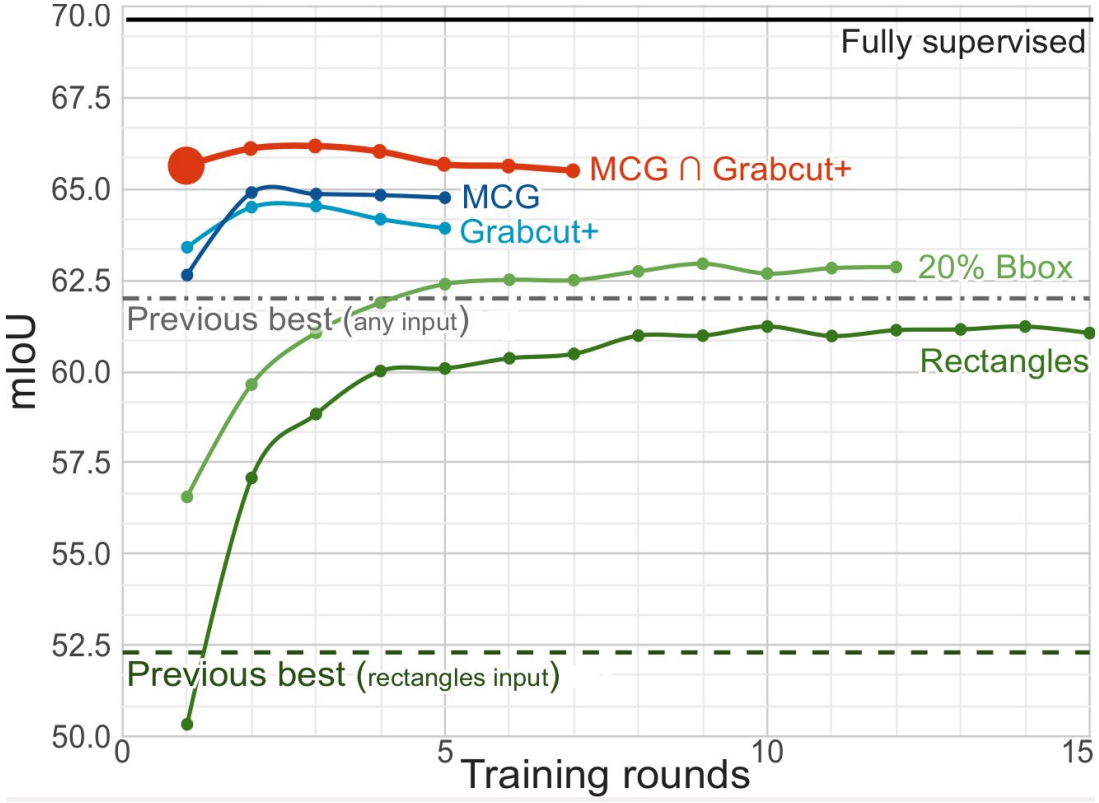




[Khoreva CVPR 17]

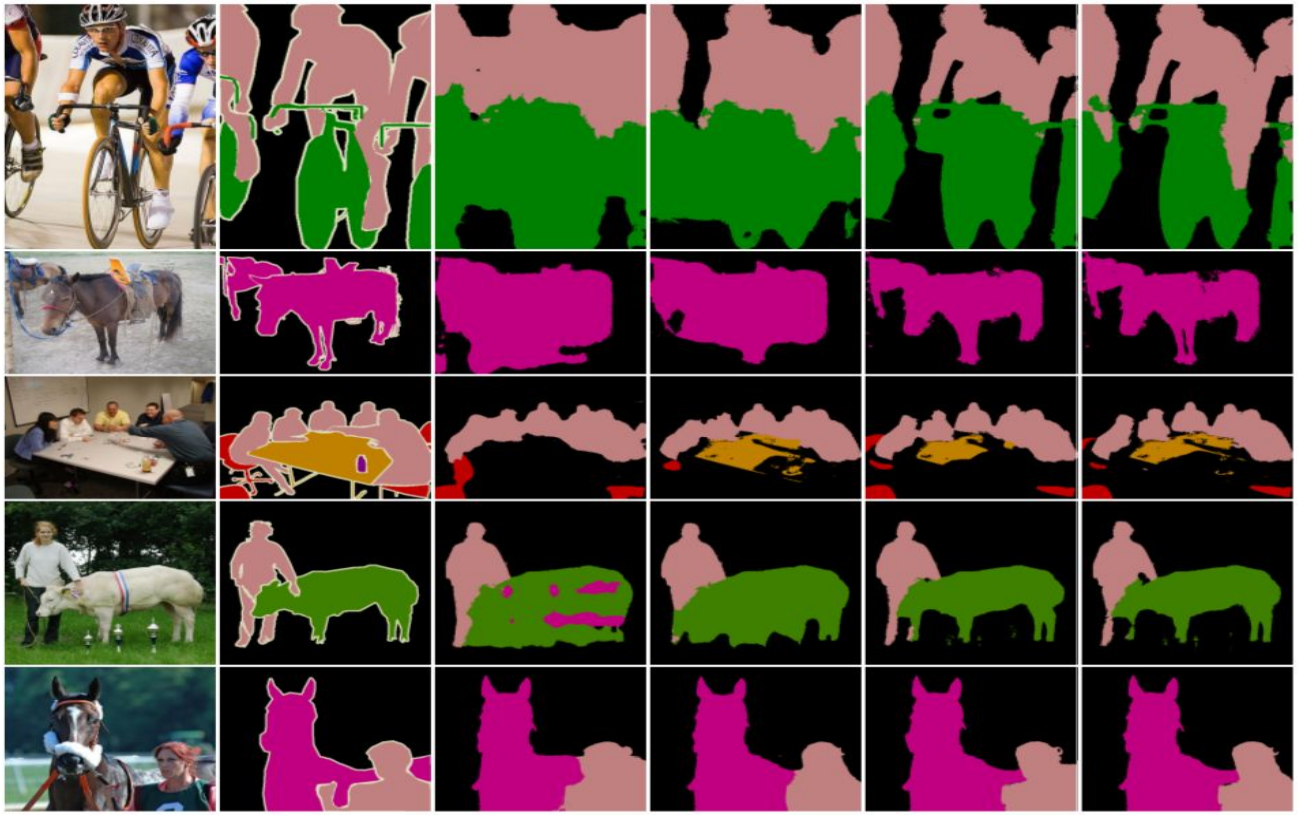
Convnets are robust to noise. Let us use that !





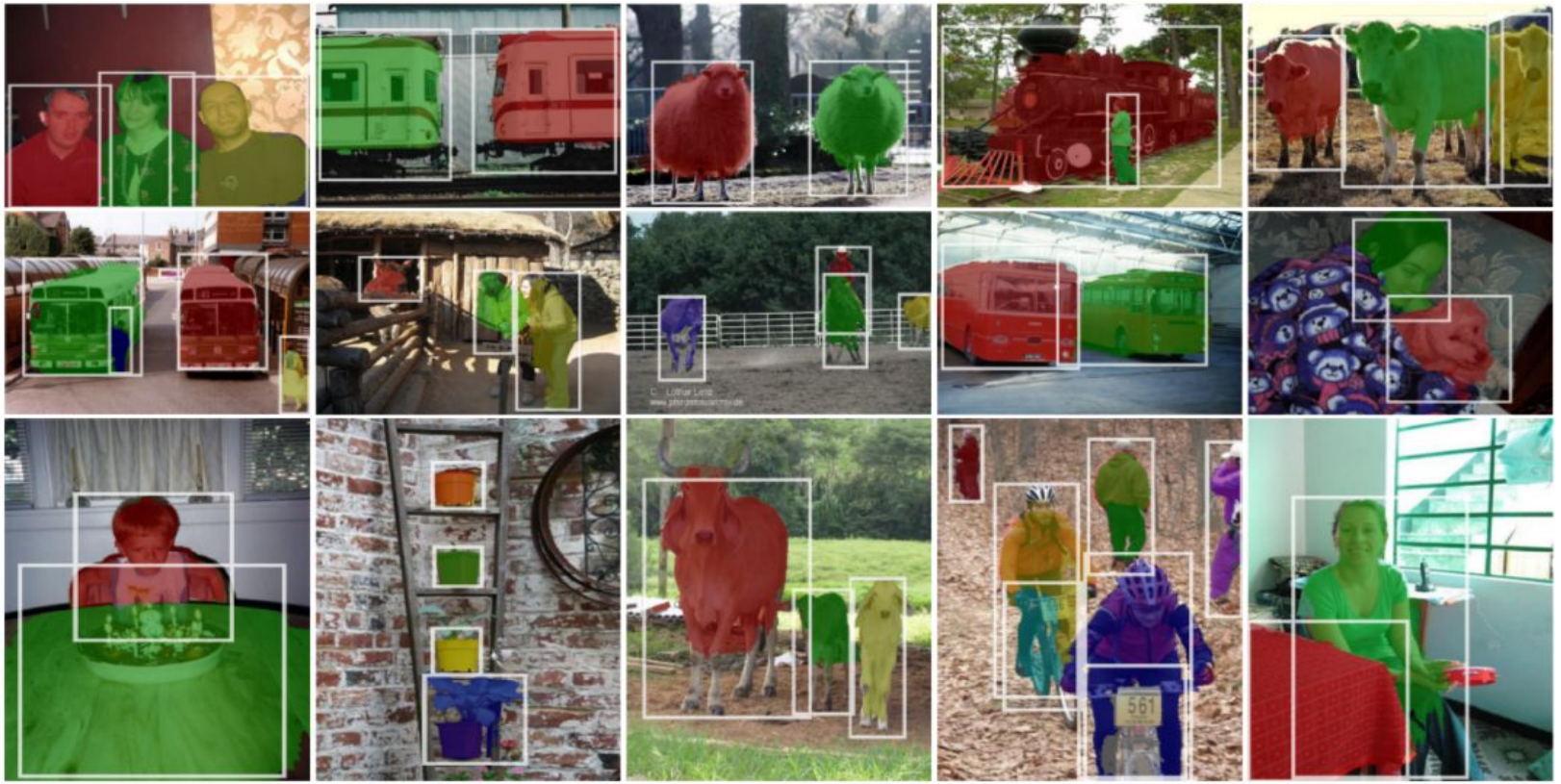
[Khoreva CVPR 17]

We reach ~95% of full supervision by using GrabCut++



[Khoreva CVPR 17]

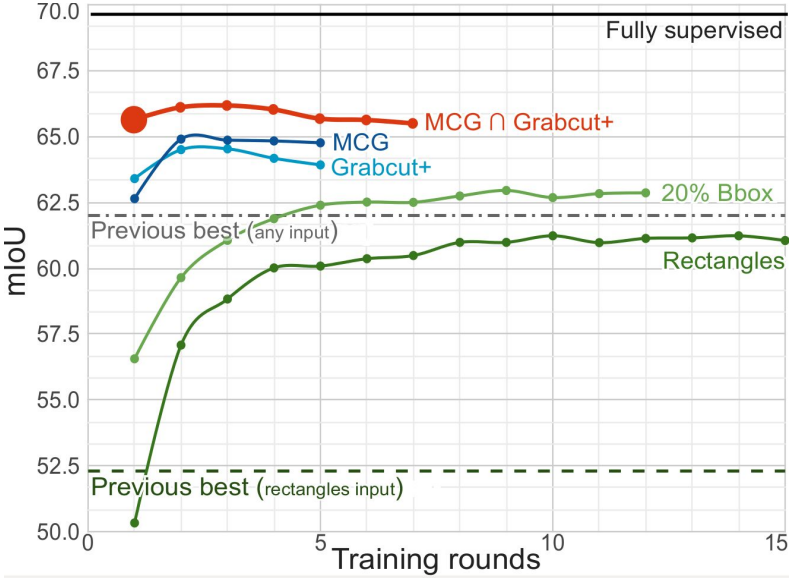
Same approach works for *instance* segmentation



[Khoreva CVPR 17]

Baselines matter !

We reach ~95% of full supervision by simply using GrabCut++



[Khoreva CVPR 17]

How else can bounding boxes be used as weak signal?

[Rother SIGGRAPH 04, Dai ICCV 15, Khoreva CVPR 17, Remez ECCV 18]

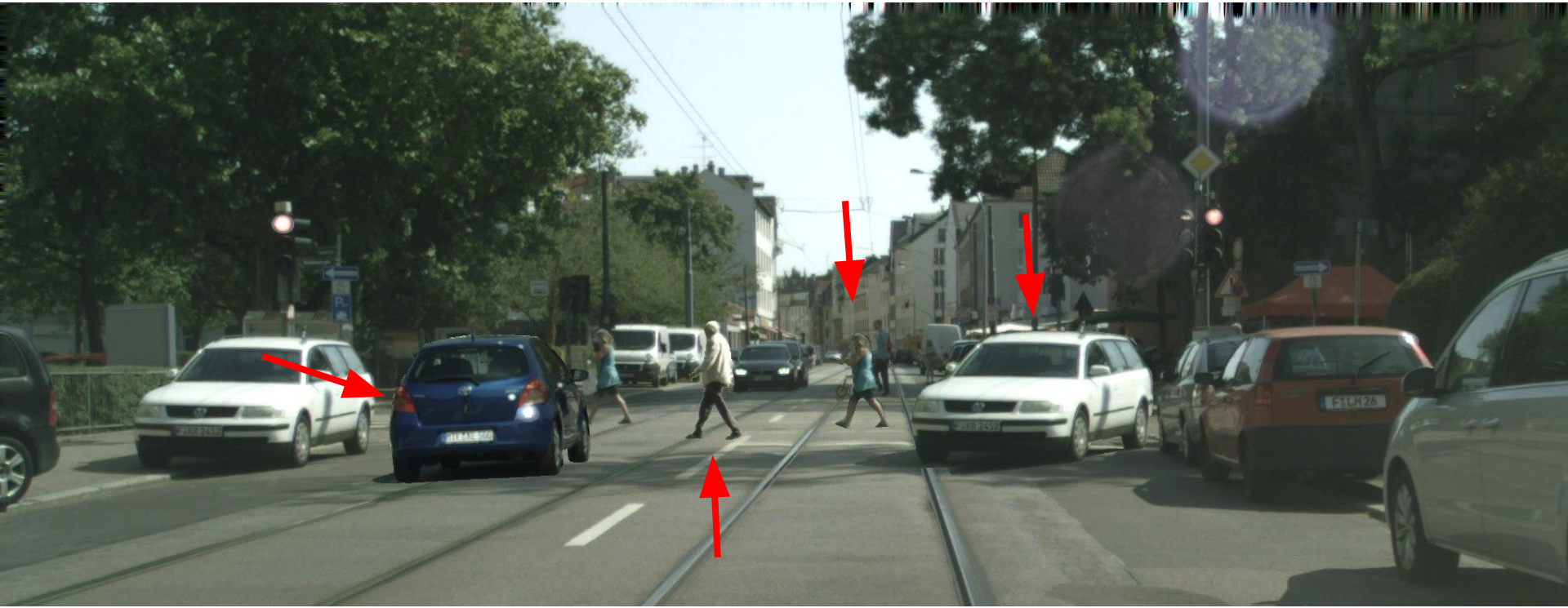
Idea:

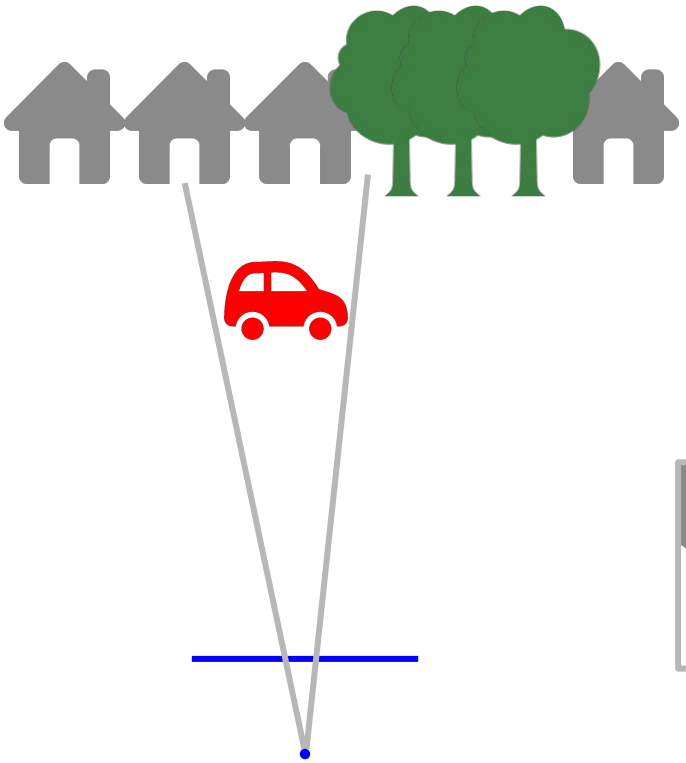
If a network can do instance segmentation, then it can cut-and-paste objects in a scene

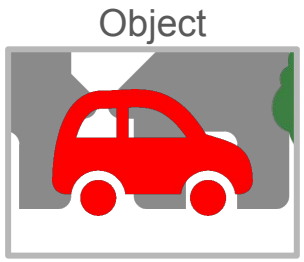
[Remez ECCV 18]

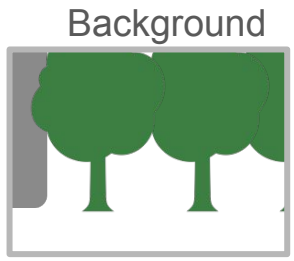
What is wrong with this image?

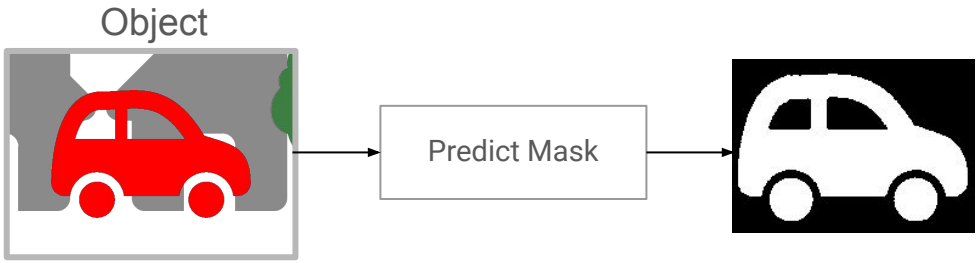
What is wrong with this image?

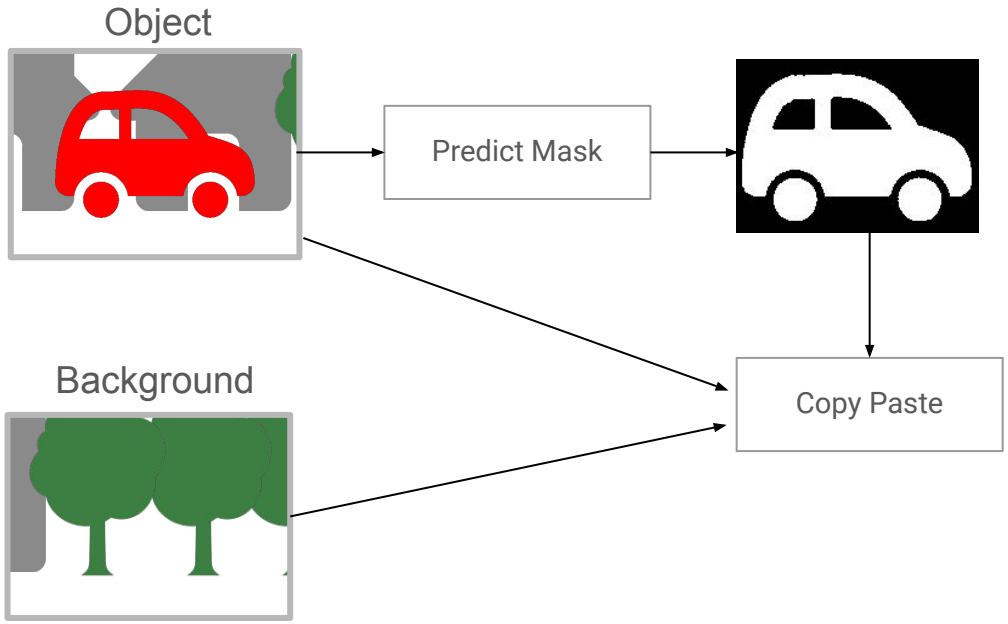


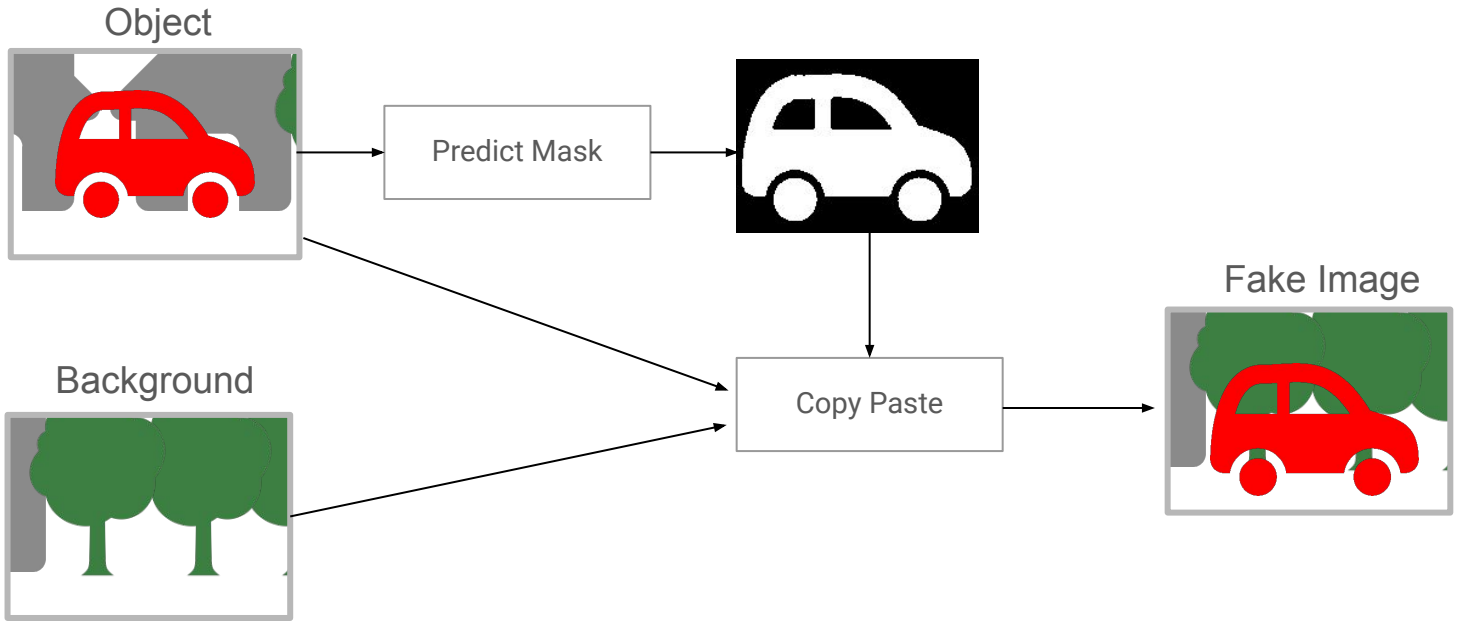


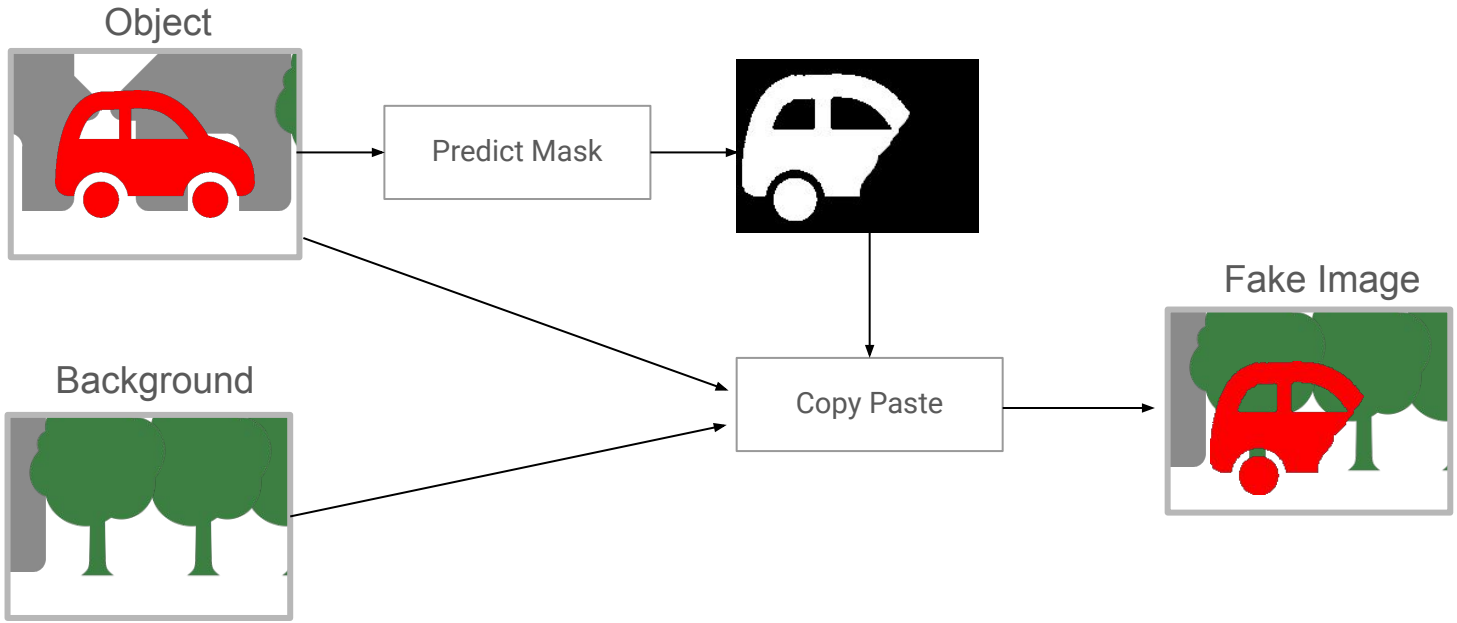

Object

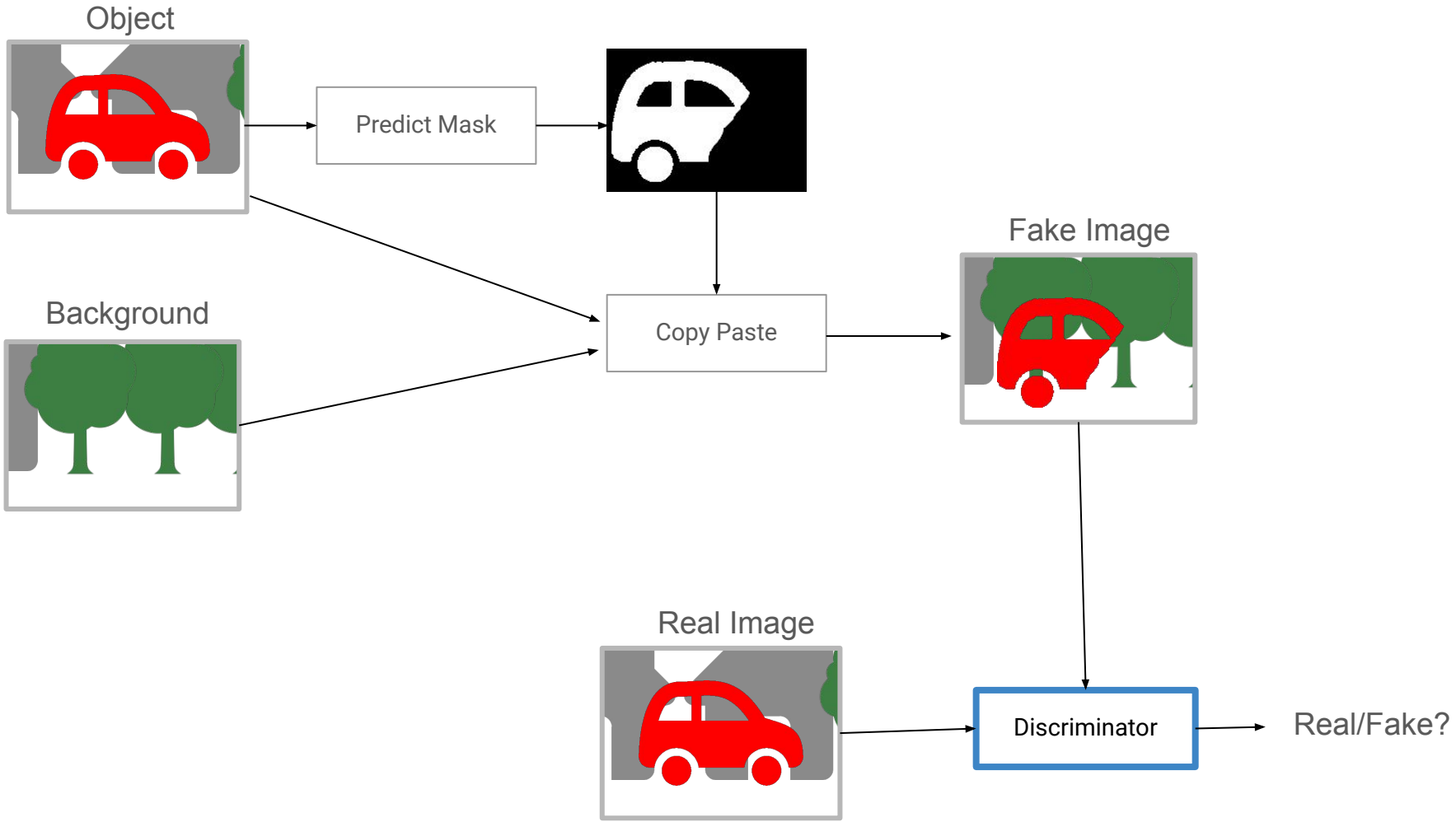

Background

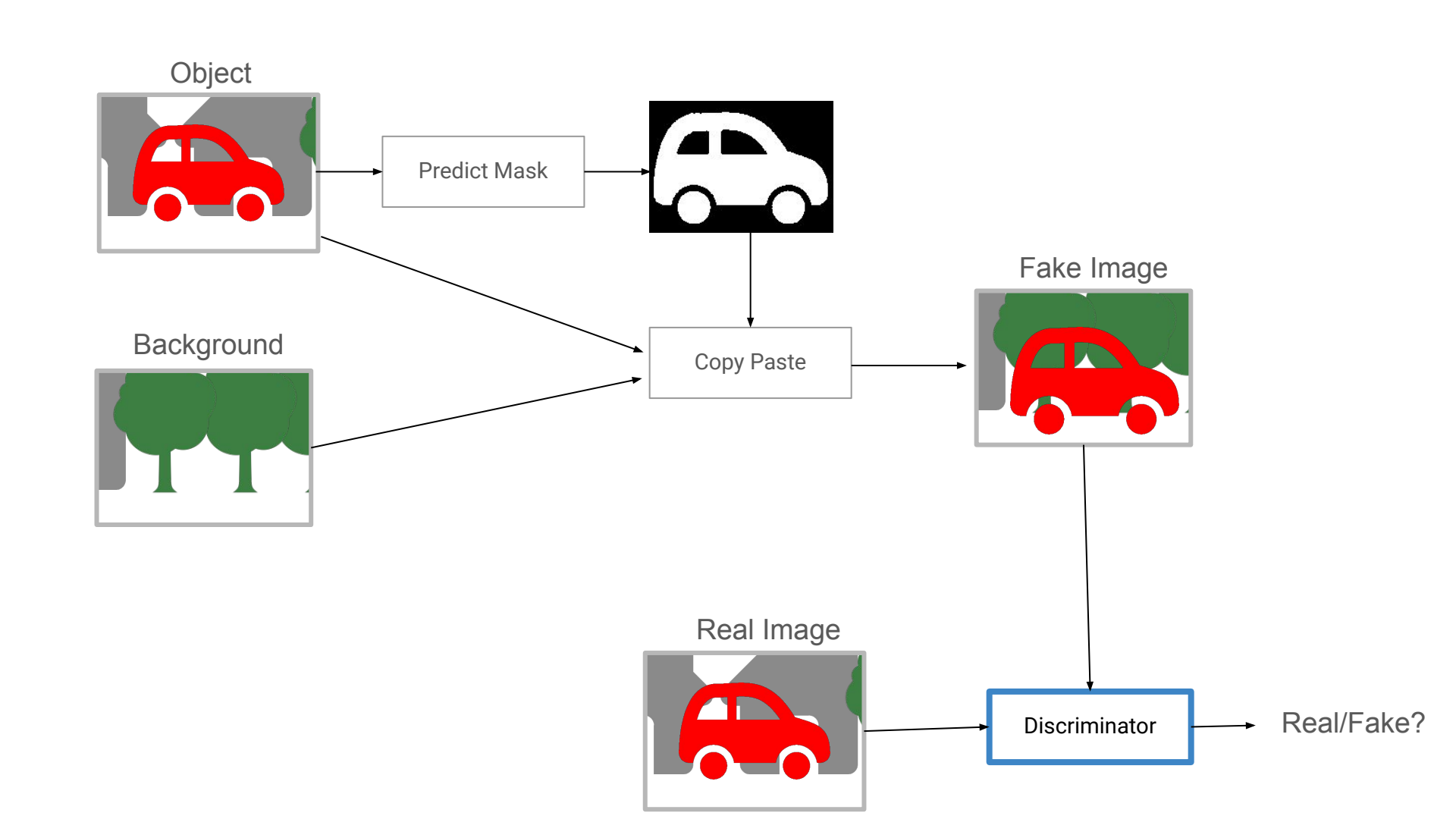

Object
Predict Mask

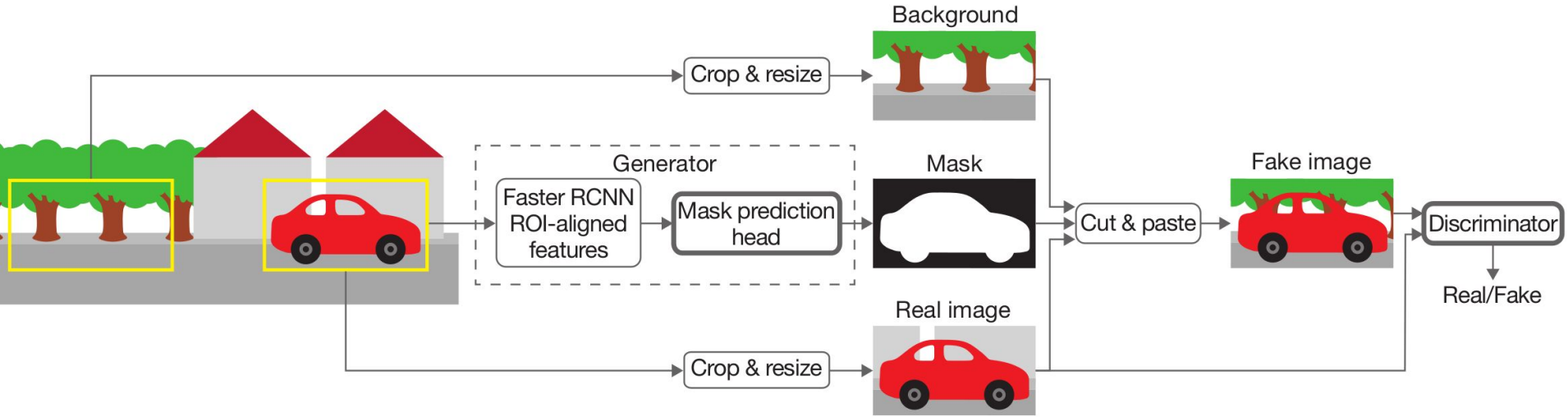

Object
Predict Mask
Background
Copy Paste


Object
Predict Mask
Background
Copy Paste
Fake Image


Object
Predict Mask
Background
Copy Paste
Fake Image


Object
Predict Mask
Background
Copy Paste
Fake Image
Real Image
Discriminator
Real/Fake?


Object
Predict Mask
Background
Copy Paste
Fake Image
Real Image
Discriminator
Real/Fake?



[Remez ECCV 18]

After careful tuning, the GAN learns !

After careful tuning, the GAN learns !

After careful tuning, the GAN learns !

Takeaways:

- Many priors and hints have been explored
- Exploit all available priors, explore new information sources

- Do not underestimate the power of transfer learning (and webly supervision)
- Current techniques are quite effective, and ride the wave of better network architectures